SMVCIR DIMENSIONALITY TEST

A Dissertation

by

CHARLES DAVID LINDSEY

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2010

Major Subject: Statistics

SMVCIR DIMENSIONALITY TEST

A Dissertation

by

CHARLES DAVID LINDSEY

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Simon J. Sheather
Committee Members,	F. Michael Speed
	Michael Longnecker
	Ronald D . Macfarlane
	Joseph R. McKean
Head of Department,	Simon J. Sheather

May 2010

Major Subject: Statistics

ABSTRACT

SMVCIR Dimensionality Test. (May 2010)
Charles Lindsey, B.A., Southwestern University;
M.S., Texas A&M University
Chair of Advisory Committee: Simon J. Sheather

The original SMVCIR algorithm was developed by Simon J. Sheather, Joseph W. McKean, and Kimberly Crimin. The dissertation first presents a new version of this algorithm that uses the scaling standardization rather than the Mahalanobis standardization. This algorithm takes grouped multivariate data as input and then outputs a new coordinate space that contrasts the groups in location, scale, and covariance. The central goal of research is to develop a method to determine the dimension of this space with statistical confidence. A dimensionality test is developed that can be used to make this determination. The new SMVCIR algorithm is compared with two other inverse regression algorithms, SAVE and SIR in the process of developing the dimensionality test and testing it.

The dimensionality test is based on the singular values of the kernel of the spanning set of the vector space. The asymptotic distribution of the spanning set is found by using the central limit theorem, delta method, and finally Slutsky's Theorem with a permutation matrix. This yields a mean adjusted asymptotic distribution of the spanning set. Theory by Eaton, Tyler, and others is then used to show an equivalence between the singular values of the mean adjusted spanning set statistic and the singular values of the spanning set statistic. The test statistic is a sample size scaled sum of squared singular values of the spanning set. This statistic is asymptotically equivalent in distribution to that of a linear combination of independent χ^2_1 random variables.

Simulations are performed to corroborate these theoretic findings. Additionally, based on work by Bentler and Xie, an approximation to the test statistic reference distribution is proposed and tested. This is also corroborated with simulations. Examples are performed that demonstrate how SMVCIR is used and how the developed tests for dimensionality are performed. Finally, further directions of research are hinted at for SMVCIR and the dimensionality test. One of the more interesting directions is explored by briefly examining how SMVCIR can be used to identify potentially complex functions that link predictors and a continuous response variable.

To my parents and all my grandparents.

ACKNOWLEDGMENTS

First of all, I would like to thank all members of my committee. I am very fortunate to have such prominent statisticians guide my work.

F. Michael Speed is best known for his research in linear models. He has also done innovative work in designing and maintaining our distance education program. He was very helpful in providing technical assistance as I completed my Master's project and my dissertation proposal.

Michael Longnecker has done great research in many areas of statistics. I was able to take my first applied statistics courses with him. As Associate Department Head, he has been a constant help as I went through the process of completing my dissertation.

Joseph R. McKean helped design the original SMVCIR algorithm. He helped me start building the current implementation of the SMVCIR algorithm in R. Dr. McKean has also provided important theoretical insight as my research continued.

Ronald D. McFarlane is a leader in Chemistry. His research problems in the Cardiovascular chemistry group at Texas A&M have inspired interest in group discrimination techniques like SMVCIR.

The Committee Chair, Simon J. Sheather is a leader in the nonparametric statistics field. I had the privilege to have him for my first course in linear regression. This experience sparked an interest in regression that I am continuing into this dissertation. Dr. Sheather's ability to organize work has allowed me to complete the dissertation in reasonable time with reasonable effort. It has also allowed us to collaborate on a variety of other intriguing projects. Sheather has provided expert guidance without mico-management, allowing me to do my best work. I would like to thank all teachers at Texas A&M Statistics. I particularly thank Jianhua Huang and Soumendra N.

Lahiri. Both taught me several interesting and very useful courses. I would also like to thank Dallas E. Johnson and Emanuel Parzen for their standout single courses. The mix of computational, theoretical, and application work I performed in my teacher's classes trained me well for the completion of my dissertation and work beyond graduation. I also thank my undergraduate instructors at Southwestern University. There I first learned to program and to perform mathematical proofs. I would thank Walter Potter especially, as he gave me my first research experience and also provided expert guidance without micro-management. Kendall C. Richards taught many of my favorite math courses and taught us a vast array of organized and useful material in each one. Suzanne Buchele taught some of my hardest and most rewarding computer science courses. She, Barbara Boucher Owens, and Richard Denman (other excellent computer science professor) taught me how to program effectively.

I would also like to thank Henrik Schmiediche, a senior lecturer in Texas A&M Statistics and the lead system administrator of our simulation systems. He also administrates the Brazos cluster, a high performance computer cluster at Texas A&M. My colleagues, at Texas A&M Statistics and Southwestern University, also deserve thanks. I have spent many hours debating fine points and cooperatively solving problems with them. We learned many things together through cooperation and disagreement. The use of this cluster was instrumental in properly performing the simulations for this dissertation.

Between undergraduate and graduate school I worked for Finis Welch at Welch Consulting. There I saw how statistics was used in the real world, and learned enough to become interested in it and come to Texas A&M. I thank Finis and my colleagues there for instilling this interest in me, and also for helping me form a good work ethic.

Finally I would like to thank my family and friends outside of school. Your support has been essential. Special thanks to my parents, for putting me through

undergrad and helping me in graduate school. Their toleration of me working late into the night and constantly hovering around our computers waiting for simulation results is very appreciated as well. All of my grandparents have also been very supportive and encouraging.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

TABLE Page

TABLE Page

TABLE Page

LIST OF FIGURES

FIGURE Page

FIGURE Page

FIGURE Page

CHAPTER I

INTRODUCTION

SMVCIR stands for Sliced Mean Variance Covariance Inverse Regression. The SMVCIR algorithm, introduced in Sheather et al. (2008), "slices" the response variable of a model into discrete groups, then discriminates between the groups using the mean, variance, and covariance differences in the predictors within each group. So we essentially regress the predictors on the response variable. We do not find differences in the response as we vary the predictors, but instead find differences in the predictors as we vary the response. This is inverse regression.

In the Sheather et al. (2008) article, a heuristic algorithm was used to select the dimensionality of the SMVCIR space, the vector space spanned by the mean, variance, and covariance differences between the response groups. In this dissertation, we will determine asymptotic distributional properties of the SMVCIR estimators. These properties will be tested in simulation. So the heuristics of the original article will be supported by well tested distribution theory.

Specifically, we will focus on one goal: the development of an algorithm that will estimate the dimension of the SMVCIR space with statistical confidence. This dimension is the number of unique mean, variance, and covariance differences between the groups. In the process of meeting this goal, several other significant tasks will performed.

We will expand the SMVCIR algorithm to utilize different predictor standardizations than the Mahalanobis. We will show how using the Mahalanobis standardization can introduce extra dimensions into the SMVCIR space representing mean, variance,

---

The journal model is *Journal of the American Statistical Association.*

and covariance differences that are not present in the non-standardized data, and offer an alternative scaling method that does not introduce new difference dimensions.

The algorithm for choosing the SMVCIR dimension will involve a series of statistical tests. The distributions of the test statistics used in these tests will be complicated. The critical points of the distributions that are used to decide which hypothesis the test suggests must be determined by simulation. Inspired by Bentler and Xie (2000), we will also develop an approximation to this distribution that will be more tractable, and allow the critical points to be determined without relying on simulation. Our theoretical work will be corroborated by simulations.

We will also provide comparison of SMVCIR with the other inverse regression algorithms, SAVE and SIR. Each of these inverse regression algorithms may make different assumptions and target different types of parameters. But they each discriminate between population groups by finding differences in mean, variance, or covariance. The simulations performed to corroborate our theoretical results will also be used to compare SMVCIR and the two other inverse regression algorithms. Following the simulations, we will have an example chapter where we will see how each of SAVE, SIR, and SMVCIR performs on datasets taken from the real world.

We evaluate SMVCIR under the assumption that the response variable has already been sliced, and that the slices are fixed. SAVE and SIR do not make this assumption, but they are still applicable when the assumption holds true. In the final chapter we examine further directions of research where we relax this assumption and allow the slicing to be data dependent.

The use of SMVCIR to estimate complex functions that link the predictors to the response variable is examined. The interface of our dimensionality test and the original SMVCIR algorithm's method for choosing the dimension is also mentioned as a further research direction. Other direction of further research involves the de-

velopment of a variable selection test for the predictors used in SMVCIR and the evaluation of the SMVCIR dimensionality test method as a multiple hypothesis test setting.

In this introductory chapter, we will describe the SMVCIR algorithm and provide an intuitive interpretation of its output. We begin with a brief example, where we show how SMVCIR works. Then we provide a more thorough discussion of SAVE and SIR. Following this we describe the general setting in which SMVCIR is invoked. Then we discuss how group means and variances should be compared. This includes an explanation of how centering can be useful and how the marginal mean and variance matrix are linked to the group means and variance matrices. We next show how standardization can be used to increase the clarity of group discrimination.

This leads us to the full definition of SMVCIR as a coordinate space and algorithm. To fully understand how this definition can be used, we first review the eigen and singular value matrix decompositions. Then with this fresh knowledge we discuss the translation of the data to the SMVCIR space, and how using the eigenvectors from the SMVCIR space's kernel illuminates the group differences.

Finally we mention the original SMVCIR ordering procedure and how it can be used to choose the dimension of the SMVCIR space.

A. Example

We will revisit this example in Chapter V and fully define the situation. For now, suppose we observe 5000 observations on six predictors from two groups. Figure 1 is a matrix plot of the data we observe. The black colored points represent group 1 while the red color points represent group 2.

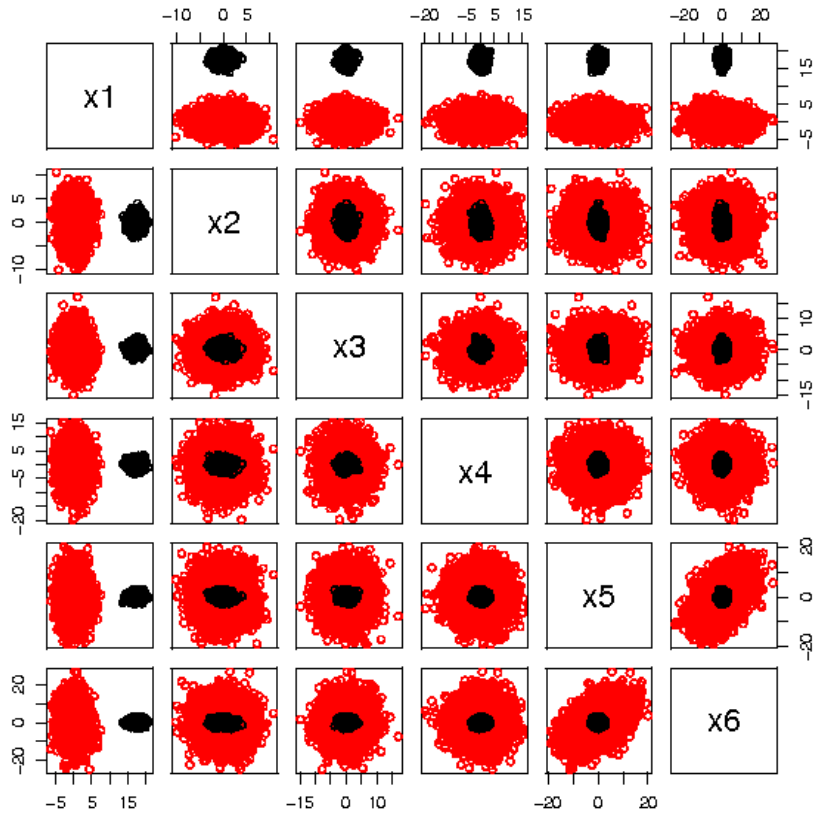

Figure 1. Ex 1.1, $n = 5000$, 2 Groups, 6 Predictors

We see obvious differences in variance between the two groups across all the predictors. A location difference is clearly present for variable x_1 across the two groups as well. The groups seem to have a different covariance relationship between x_5 and x_6 also.

It can be somewhat cumbersome to identify all these differences by visual examination of the matrix plot of the predictors. Moreover, we cannot attach statistical confidence to our identifications.

The SMVCIR algorithm takes the data we just observed, and after running the

dimensionality test gives the following transformed coordinates in Figure 2.

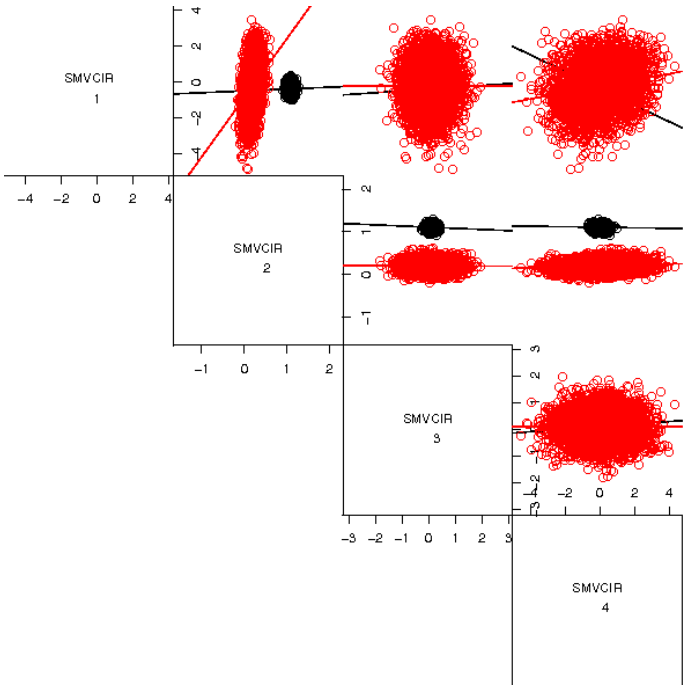

Figure 2. SMVCIR Coordinates, Ex. 1.1

These coordinates show the differences between the groups in a more succinct fashion than the original scatter plot (Figure 1). The first coordinate represents the variance difference. The second represents the location (mean) difference. The third and fourth represent the covariance difference. We will discuss why we need two dimensions to represent a single covariance difference later in this chapter.

The covariance differences are more easily seen in a three dimensional plot of Figure 3. The right plot shows the differing slopes between the two groups (black and red) very clearly.

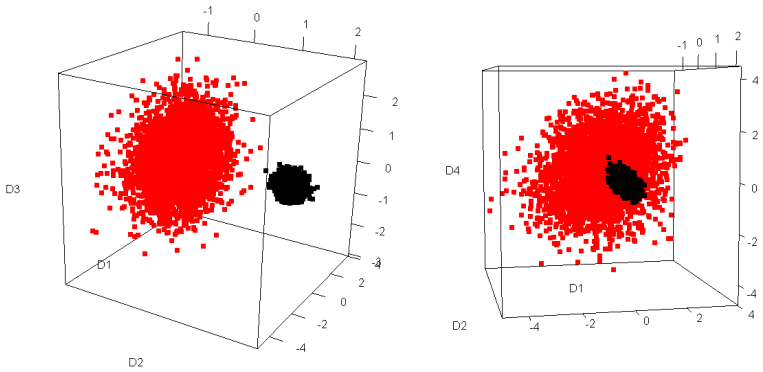

Figure 3. 3D SMVCIR Coordinates, Ex. 1.1

So SMVCIR takes as input the original predictor data, and then outputs transformed data that more clearly and concisely shows the mean, variance, and covariance differences between the groups. We showed that it may take fewer than $d \leq k$ (k the number of predictors) transformed coordinates to show this. This quantity d is the dimension of the SMVCIR space, and it is the primary goal of the dissertation to develop a method to find d with statistical confidence.

B. SAVE and SIR

There are other inverse regression algorithms. SIR, Sliced Inverse Regression was proposed in Li (1991). This algorithm targets the location differences between response groups. SAVE, Sliced Average Variance Estimation was presented in Cook (2000). This algorithm targets both location and variance/covariance differences between the

response groups. Moreover, it seeks to capture the central subspace of the regression of the response on the predictors. The central subspace is the smallest vector space formed by a projection of the predictors S, such that the response is independent of the predictors conditional on S.

The SMVCIR method is novel in the way it handles variance differences. SAVE treats variance and covariance differences in identical manners, so when two groups differ in variance but not covariance for six predictors, SAVE will find six dimensions for that difference. SMVCIR stacks the variance differences together separately from covariance differences, so it would find only one dimension for the six variance differences.

The advantages and disadvantages of SMVCIR versus each of these algorithms are discussed in the original Sheather et al. (2008) paper. Some new insights on the comparison of the three procedures will be gained in this project, and we will highlight these as we proceed.

C. Setting

In a situation where we may invoke SMVCIR, we observe n independent and identically distributed $(k+1) \times 1$ random variates, $(\mathbf{x_1'}, y_1)', \ldots, (\mathbf{x_n'}, y_n)'$. We refer to the k variables that begin each observation as our predictor variables, and the final variable of each observation as our response variable. This ordering is performed so that our theoretical results are more easily derived. It should be a simple task to re-order the variables to meet the criteria we have given in any applied situation.

The original SMVCIR algorithm was designed for continuous predictors. Our results for the dimensionality and variable selection tests should apply to continuous or discrete predictors. For brevity, we will restrict our attention to continuous predictors

here, and leave examination of discrete predictors for later work. The response can be discrete, continuous or from mixture distributions. If the response is not discrete, we will make it discrete by slicing. Slicing partitions the support of the input variable based on the signed magnitude of the input variable's value. We require that the range of the slices be defined before data is sampled. Values of the response that are close together will be put into the same group. Once the response is discretized, its unique discrete values are assigned numeric codes $1, \ldots, g$. This number indicates what group we place the response into. Since we require that the slices on the response be fixed before sampling, we fix the discretized group definition of y before sampling.

We also require that the expectation vector and covariance matrix are defined (non-infinite) for the predictors, marginally and conditioned on the group recoded value of the response. Moreover, the marginal and conditioned covariance matrices must be non-singular.

In SMVCIR, we wish to contrast the means, variances, and covariances between different groups. We do this by estimating differences between the means, variances, and covariances in each group and the overall average means, variances, and covariances. This is more easily explained in the context of the population. The predictors in response group i have mean $\boldsymbol{\mu}_\mathbf{i}$ and variance matrix $\boldsymbol{\Sigma}_\mathbf{i}$. The probability of drawing an observation from group i when randomly sampling from the population is p_i.

D. Group Mean and Variance/Covariance Comparison

Let us calculate the marginal mean vector $\boldsymbol{\mu}$, ignoring the grouping index variable, which we will call y. We call our predictor vector $\mathbf{x}$.

$$\begin{aligned}
\boldsymbol{\mu} &= \mathrm{E}\left[\mathbf{x}\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\mathbf{x}|y\right]\right] \\
&= \mathrm{E}\left[\sum_{i=1}^{g} \boldsymbol{\mu}_{\mathbf{i}} \mathrm{I}\left(y=i\right)\right] \\
&= \sum_{i=1}^{g} \boldsymbol{\mu}_{\mathbf{i}} \mathrm{E}\left[\mathrm{I}\left(y=i\right)\right] \\
&= \sum_{i=1}^{g} p_i \boldsymbol{\mu}_{\mathbf{i}}
\end{aligned} \tag{1.1}$$

Now we will compute the marginal variance matrix.

$$\begin{aligned}
\boldsymbol{\Sigma} &= \mathrm{Var}\left[\mathbf{x}\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\left(\mathbf{x}-\boldsymbol{\mu}\right)\left(\mathbf{x}-\boldsymbol{\mu}\right)'|y\right]\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\mathbf{x}\mathbf{x}'-\boldsymbol{\mu}\mathbf{x}'-\mathbf{x}\boldsymbol{\mu}'+\boldsymbol{\mu}\boldsymbol{\mu}'|y\right]\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\mathbf{x}\mathbf{x}'|y\right]-\mathrm{E}\left[\boldsymbol{\mu}\mathbf{x}'|y\right]-\mathrm{E}\left[\mathbf{x}\boldsymbol{\mu}'|y\right]+\boldsymbol{\mu}\boldsymbol{\mu}'\right] \\
&= \mathrm{E}\left[\mathrm{E}\left[\mathbf{x}\mathbf{x}'|y\right]\right]-\mathrm{E}\left[\mathrm{E}\left[\boldsymbol{\mu}\mathbf{x}'|y\right]\right]-\mathrm{E}\left[\mathrm{E}\left[\mathbf{x}\boldsymbol{\mu}'|y\right]\right]+\boldsymbol{\mu}\boldsymbol{\mu}'
\end{aligned} \tag{1.2}$$

Next we compute the individual expectations. We begin with $\mathrm{E}\left[\mathrm{E}\left[\mathbf{x}\mathbf{x}'|y\right]\right]$.

$$\begin{aligned}
\mathrm{E}\left[\mathrm{E}\left[\mathbf{x}\mathbf{x}'|y\right]\right] &= \mathrm{E}\left[\mathrm{Var}\left[\mathbf{x}|y\right]+\mathrm{E}\left[\mathbf{x}|y\right]\mathrm{E}\left[\mathbf{x}|y\right]'\right] \\
&= \mathrm{E}\left[\sum_{i=1}^{g} \boldsymbol{\Sigma}_{\mathbf{i}} \mathrm{I}\left(y=i\right)+\left(\sum_{i=1}^{g} \boldsymbol{\mu}_{\mathbf{i}} \mathrm{I}\left(y=i\right)\right)\left(\sum_{i=1}^{g} \boldsymbol{\mu}_{\mathbf{i}} \mathrm{I}\left(y=i\right)\right)'\right] \\
&= \mathrm{E}\left[\sum_{i=1}^{g} \boldsymbol{\Sigma}_{\mathbf{i}} \mathrm{I}\left(y=i\right)+\sum_{i=1}^{g}\sum_{j=1}^{g} \boldsymbol{\mu}_{\mathbf{i}}\boldsymbol{\mu}_{\mathbf{j}}' \mathrm{I}\left(y=i\right)\mathrm{I}\left(y=j\right)\right] \\
&= \sum_{i=1}^{g} \boldsymbol{\Sigma}_{\mathbf{i}} \mathrm{E}\left[\mathrm{I}\left(y=i\right)\right]+\sum_{i=1}^{g}\sum_{j=1}^{g} \boldsymbol{\mu}_{\mathbf{i}}\boldsymbol{\mu}_{\mathbf{j}}' \mathrm{E}\left[\mathrm{I}\left(y=i\right)\mathrm{I}\left(y=j\right)\right] \\
&= \sum_{i=1}^{g} \boldsymbol{\Sigma}_{\mathbf{i}} p_i+\sum_{i=1}^{g}\sum_{j=1}^{g} \boldsymbol{\mu}_{\mathbf{i}}\boldsymbol{\mu}_{\mathbf{j}}' \left(\mathrm{I}\left(j=i\right)p_i\right) \\
&= \sum_{i=1}^{g} \boldsymbol{\Sigma}_{\mathbf{i}} p_i+\sum_{j=1}^{g} \boldsymbol{\mu}_{\mathbf{i}}\boldsymbol{\mu}_{\mathbf{i}}' p_i
\end{aligned} \tag{1.3}$$

$$\begin{aligned}
\mathrm{E}\left[\mathrm{E}\left[\boldsymbol{\mu}\mathbf{x}'|y\right]\right] &= \mathrm{E}\left[\boldsymbol{\mu}\mathrm{E}\left[\mathbf{x}'|y\right]\right] \\
&= \mathrm{E}\left[\sum_{i=1}^{g}\boldsymbol{\mu}\boldsymbol{\mu}_{\mathbf{i}}'\mathrm{I}\left(y=i\right)\right] \\
&= \sum_{i=1}^{g}\boldsymbol{\mu}\boldsymbol{\mu}_{\mathbf{i}}'\mathrm{E}\left[\mathrm{I}\left(y=i\right)\right] \\
&= \sum_{i=1}^{g}\boldsymbol{\mu}\boldsymbol{\mu}_{\mathbf{i}}'p_i \\
&= \boldsymbol{\mu}\sum_{i=1}^{g}\boldsymbol{\mu}_{\mathbf{i}}'p_i \\
&= \boldsymbol{\mu}\boldsymbol{\mu}'
\end{aligned} \tag{1.4}$$

$$\begin{aligned}
\mathrm{E}\left[\mathrm{E}\left[\mathbf{x}\boldsymbol{\mu}'|y\right]\right] &= \mathrm{E}\left[\mathrm{E}\left[\mathbf{x}|y\right]\boldsymbol{\mu}'\right] \\
&= \mathrm{E}\left[\sum_{i=1}^{g}\boldsymbol{\mu}_{\mathbf{i}}\boldsymbol{\mu}'\mathrm{I}\left(y=i\right)\right] \\
&= \sum_{i=1}^{g}\boldsymbol{\mu}_{\mathbf{i}}\boldsymbol{\mu}'\mathrm{E}\left[\mathrm{I}\left(y=i\right)\right] \\
&= \sum_{i=1}^{g}\boldsymbol{\mu}_{\mathbf{i}}\boldsymbol{\mu}'p_i \\
&= \left(\sum_{i=1}^{g}\boldsymbol{\mu}_{\mathbf{i}}p_i\right)\boldsymbol{\mu}' \\
&= \boldsymbol{\mu}\boldsymbol{\mu}'
\end{aligned} \tag{1.5}$$

Now we will apply (1.3)-(1.5) to our calculation of $\boldsymbol{\Sigma}$.

$$\begin{aligned}
\boldsymbol{\Sigma} &= \mathrm{Var}\left[\mathbf{x}\right] \\
&= \left(\sum_{i=1}^{g}\boldsymbol{\Sigma}_{\mathbf{i}}p_i + \sum_{j=1}^{g}\boldsymbol{\mu}_{\mathbf{i}}\boldsymbol{\mu}_{\mathbf{i}}'p_i\right) - \boldsymbol{\mu}\boldsymbol{\mu}' - \boldsymbol{\mu}\boldsymbol{\mu}' + \boldsymbol{\mu}\boldsymbol{\mu}' \\
&= \sum_{i=1}^{g}\boldsymbol{\Sigma}_{\mathbf{i}}p_i + \sum_{j=1}^{g}\boldsymbol{\mu}_{\mathbf{i}}\boldsymbol{\mu}_{\mathbf{i}}'p_i - \boldsymbol{\mu}\boldsymbol{\mu}'
\end{aligned} \tag{1.6}$$

As in the univariate case, we can partition the marginal variance into the expected conditional variance and the variance of the conditional expectation.

$$\begin{aligned}
\mathrm{E}\left[\mathrm{Var}\left[\mathbf{x}|y\right]\right] &= \mathrm{E}\left[\sum_{i=1}^{g}\boldsymbol{\Sigma}_{\mathbf{i}}\mathrm{I}\left(y=i\right)\right] \\
&= \sum_{i=1}^{g}\boldsymbol{\Sigma}_{\mathbf{i}}\mathrm{E}\left[\mathrm{I}\left(y=i\right)\right] \\
&= \textstyle\sum_{i=1}^{g}p_i\boldsymbol{\Sigma}_{\mathbf{i}}
\end{aligned} \tag{1.7}$$

$$
\begin{aligned}
\mathrm{Var}\left[\mathrm{E}\left[\mathbf{x}|y\right]\right] &= \mathrm{Var}\left[\sum_{i=1}^{g} \boldsymbol{\mu}_{\mathbf{i}} \mathrm{I}\left(y=i\right)\right] \\
&= \mathrm{E}\left[\left(\sum_{i=1}^{g} \boldsymbol{\mu}_{\mathbf{i}} \mathrm{I}\left(y=i\right)\right)\left(\sum_{i=1}^{g} \boldsymbol{\mu}_{\mathbf{i}} \mathrm{I}\left(y=i\right)\right)'\right] - \\
&\quad \mathrm{E}\left[\sum_{i=1}^{g} \boldsymbol{\mu}_{\mathbf{i}} \mathrm{I}\left(y=i\right)\right] \mathrm{E}\left[\sum_{i=1}^{g} \boldsymbol{\mu}_{\mathbf{i}} \mathrm{I}\left(y=i\right)\right]' \\
&= \mathrm{E}\left[\sum_{i=1}^{g}\sum_{i=1}^{g} \boldsymbol{\mu}_{\mathbf{i}}\boldsymbol{\mu}_{\mathbf{j}}' \mathrm{I}\left(y=i\right)\mathrm{I}\left(y=j\right)\right] - \\
&\quad \left(\sum_{i=1}^{g} p_i \boldsymbol{\mu}_{\mathbf{i}}\right)\left(\sum_{i=1}^{g} p_i \boldsymbol{\mu}_{\mathbf{i}}\right)' \\
&= \mathrm{E}\left[\sum_{i=1}^{g} \boldsymbol{\mu}_{\mathbf{i}}\boldsymbol{\mu}_{\mathbf{i}}' \mathrm{I}\left(y=i\right)\right] - \\
&\quad \left(\sum_{i=1}^{g} p_i \boldsymbol{\mu}_{\mathbf{i}}\right)\left(\sum_{i=1}^{g} p_i \boldsymbol{\mu}_{\mathbf{i}}\right)' \\
&= \sum_{j=1}^{g} \boldsymbol{\mu}_{\mathbf{i}}\boldsymbol{\mu}_{\mathbf{i}}' p_i - \boldsymbol{\mu}\boldsymbol{\mu}'
\end{aligned}
\tag{1.8}
$$

Let $\boldsymbol{\mu}_{\mathbf{i}}^{\mathbf{c}} = \boldsymbol{\mu}_{\mathbf{i}} - \boldsymbol{\mu}$ be the centered un-standardized means. Compare the following two matrices of group means.

$$
\begin{bmatrix}
\boldsymbol{\mu}_{\mathbf{1}} & \cdots & \boldsymbol{\mu}_{\mathbf{g}} \\
\boldsymbol{\mu}_{\mathbf{1}}^{\mathbf{c}} & \cdots & \boldsymbol{\mu}_{\mathbf{g}}^{\mathbf{c}}
\end{bmatrix}
\tag{1.9}
$$

The column i in the first matrix shares the same rank and difference relationships with the other columns of the first matrix, as column i of the second matrix shares with the other columns of the second matrix. By rank and difference relationships, we mean properties under the binary operators $=, <, >$.

So both the centered and un-centered means provide the same information about the presence of location differences between groups. However, a single centered mean $\boldsymbol{\mu}_{\mathbf{i}}^{\mathbf{c}}$ provides more information about differences than the corresponding un-centered mean $\boldsymbol{\mu}_{\mathbf{i}}$. The un-centered mean $\boldsymbol{\mu}_{\mathbf{i}}$ shows the distance of the group i in location from the origin, or the absolute center $\mathbf{0}$. The centered mean shows the distance of

the group i in location from the relative center of the predictor data, $\boldsymbol{\mu}$. We can tell whether a predictor component of the group i is typically above or below the average value of that predictor over the entire data by using $\boldsymbol{\mu}_{\mathbf{i}}^{\mathbf{c}}$.

We define the centered variances similarly, $\boldsymbol{\Sigma}_{\mathbf{i}}^{\mathbf{c}} = \boldsymbol{\Sigma}_{\mathbf{i}} - \sum_{i=1}^{g} p_i \boldsymbol{\Sigma}_{\mathbf{i}}$. Note how $\boldsymbol{\Sigma}$ is not used. The pooled mean is identical to the marginal mean. This is of course not the case for the variances. We only use the pooled variance (expected conditional variance) $\mathrm{E}\left[\mathrm{Var}\left[\mathbf{x}|y\right]\right]$ to center, and ignore $\mathrm{Var}\left[\mathrm{E}\left[\mathbf{x}|y\right]\right]$.

There is a direct correspondence between the pooled variance and actual variance when the means are equal. In this situation, we can interpret the elements of $\boldsymbol{\Sigma}_{\mathbf{i}}^{\mathbf{c}}$ as we did the elements of $\boldsymbol{\mu}_{\mathbf{i}}^{\mathbf{c}}$, telling us about the scaling of the group i rather than its location relative to the marginal scaling of the groups.

When the means are different, the pooled variance is no longer an accurate measure of the marginal variance. So the centered variances are not centered with regard to the actual central scaling of the predictor population. Are we correct to center using the pooled variance in this case?

There are cases where the group means will greatly differ and have a large effect on $\boldsymbol{\Sigma}$. Suppose we have two groups and three predictors.

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{1}} &= -\begin{bmatrix} 100 \\ 100 \\ 75 \end{bmatrix} \quad \boldsymbol{\mu}_{\mathbf{2}} = \begin{bmatrix} 100 \\ 100 \\ 75 \end{bmatrix} \\ p_1 &= p_2 = .5 \\ \boldsymbol{\Sigma}_{\mathbf{1}} &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \boldsymbol{\Sigma}_{\mathbf{2}} = \begin{bmatrix} 1.2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{aligned} \tag{1.10}$$

The expected conditional variance is drastically different from the marginal vari-

ance.

$$\mathrm{E}\left[\mathrm{Var}\left[\mathbf{x}|y\right]\right] = \begin{bmatrix} .5 + .5 * 1.2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1.1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{1.11}$$

$$\begin{aligned} &\mathrm{Var}\left[\mathrm{E}\left[\mathbf{x}|y\right]\right] = \\ &\sum_{j=1}^{2} \boldsymbol{\mu_i}\boldsymbol{\mu_i'}p_i - \boldsymbol{\mu}\boldsymbol{\mu'} = \\ &.5 \begin{bmatrix} -100 \\ -100 \\ -75 \end{bmatrix} \begin{bmatrix} -100 \\ -100 \\ -75 \end{bmatrix}' + .5 \begin{bmatrix} 100 \\ 100 \\ 75 \end{bmatrix} \begin{bmatrix} 100 \\ 100 \\ 75 \end{bmatrix}' - 0 = \\ &\begin{bmatrix} 100 \\ 100 \\ 75 \end{bmatrix} \begin{bmatrix} 100 \\ 100 \\ 75 \end{bmatrix}' = \begin{bmatrix} 10000 & 10000 & 7500 \\ 10000 & 10000 & 7500 \\ 7500 & 7500 & 5625 \end{bmatrix} \end{aligned} \tag{1.12}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} 10001.1 & 10000 & 7500 \\ 10000 & 10001 & 7500 \\ 7500 & 7500 & 5626 \end{bmatrix} \tag{1.13}$$

The marginal variance is dominated by the location differences between the groups. When we use $\boldsymbol{\Sigma}_\mathbf{i}^\mathbf{c}$ we are interested in only variance differences, so using a center measure that can be dominated by mean differences seems awkward.

So, in summary, we interpret the elements of $\boldsymbol{\Sigma}_\mathbf{i}^\mathbf{c}$ as telling about the scaling of the group i relative to the average (not necessarily marginal) scaling of the groups. The ranking relationships between the group variances remain invariant between the un-centered and centered group variances as well. We elucidated this type of invariance with the centered and un-centered means by arraying them in (1.9).

E. Standardization

The SMVCIR algorithm works on the centered parameter vectors and matrices, detecting deviations from the center through these difference vectors and matrices. We

do not directly use the centered means and variances. Two factors can obscure the difference information in the original centered means and variances: marginal linear dependency and scaling.

SMVCIR will translate the original data into new coordinates that show the mean, variance, and covariance differences between the response groups. When there is marginal linear dependency among the variables, they are correlated in $\mathbf{\Sigma}$ and they may remain correlated under the translation, so visualization and interpretation of the new coordinates may be difficult. When the marginal scaling drastically differs in $\mathbf{\Sigma}$ for different predictors, their transformed scales may drastically differ as well. Visualization may be easier when coordinates share similar scales.

To correct these problems, we introduce a standardization matrix $\mathbf{T_z}$ and standardize the centered means and centered variances. This operation can be interpreted as standardizing the marginal predictor populations. Then the centered means and variances are recomputed from the standardized marginal population. Since matrix multiplication distributes over matrix addition, we have the following.

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{i}}^{\mathbf{cz}} &= \mathbf{T_z}\boldsymbol{\mu}_{\mathbf{i}}^{\mathbf{c}} \\ \mathbf{\Sigma}_{\mathbf{i}}^{\mathbf{cz}} &= \mathbf{T_z}\mathbf{\Sigma}_{\mathbf{i}}^{\mathbf{c}}\mathbf{T}_{\mathbf{z}}' \end{aligned} \tag{1.14}$$

We will discover that correcting marginal dependency using a non-diagonal matrix for $\mathbf{T_z}$ may violate invariance under the $=, >, <$ operators for components of the transformed differences. So the relationship between components of $\boldsymbol{\mu}_{\mathbf{j}}^{\mathbf{cz}}$ and $\boldsymbol{\mu}_{\mathbf{i}}^{\mathbf{cz}}$ under $=, >, <$ may differ from that of the same components of $\boldsymbol{\mu}_{\mathbf{j}}^{\mathbf{c}}$ and $\boldsymbol{\mu}_{\mathbf{i}}^{\mathbf{c}}$. Similarly, the relationship between components of $\mathbf{\Sigma}_{\mathbf{j}}^{\mathbf{cz}}$ and $\mathbf{\Sigma}_{\mathbf{i}}^{\mathbf{cz}}$ under $=, >, <$ may differ from that of the same components of $\mathbf{\Sigma}_{\mathbf{j}}^{\mathbf{c}}$ and $\mathbf{\Sigma}_{\mathbf{i}}^{\mathbf{c}}$. But using a diagonal matrix that only corrects marginal scaling problems may lead to unacceptable correlations in the transformed coordinates. These complications will be addressed in the next chapter.

Note that we can use the superscripts cz and zc interchangeably, due to the distribution of multiplication over addition. We will use whichever subscript best reflects the order of operations we used to standardize and center (or center and standardize).

F. SMVCIR Space and Spanning Set

Let the $\boldsymbol{\sigma}_{\mathbf{j}}^{\mathbf{cz}} = \text{diag}\left(\boldsymbol{\Sigma}_{\mathbf{j}}^{\mathbf{cz}}\right)$ vectors represent the standardized centered variances (diag is short for diagonal). The $\breve{\boldsymbol{\Sigma}}_{\mathbf{j}}^{\mathbf{cz}} = \boldsymbol{\Sigma}_{\mathbf{j}}^{\mathbf{cz}} - \text{diag}\left(\boldsymbol{\sigma}_{\mathbf{j}}^{\mathbf{cz}}\right)$ matrices represent the standardized centered covariance matrices, with the variance components zeroed out. This variance stacking procedure provides an efficient way to represent variance differences. When there are multiple variance differences and no covariance differences, only one unique vector is needed to show the difference. Without the stacking, a separate vector would be necessary for each component difference.

The SMVCIR space is the column space of the following matrix.

$$\mathbf{K} = \left[\sqrt{p_1}\boldsymbol{\mu}_{\mathbf{1}}^{\mathbf{cz}} \;\; \cdots \;\; \sqrt{p_g}\boldsymbol{\mu}_{\mathbf{g}}^{\mathbf{cz}} \;\; \sqrt{p_1}\sigma_{\mathbf{1}}^{\mathbf{cz}} \;\; \cdots \;\; \sqrt{p_g}\sigma_{\mathbf{g}}^{\mathbf{cz}} \;\; \sqrt{p_1}\breve{\boldsymbol{\Sigma}}_{\mathbf{1}}^{\mathbf{cz}} \;\; \cdots \sqrt{p_g}\breve{\boldsymbol{\Sigma}}_{\mathbf{g}}^{\mathbf{cz}} \right] \tag{1.15}$$

So each difference is a dimension of the SMVCIR space, which we denote $\boldsymbol{\Omega}_{\mathbf{SMVCIR}}$. We scale the spanning vectors corresponding with each group by the square root of that group's population proportion. Each spanning set vector represents the difference of a particular group from the overall group average. This weighting gives a greater magnitude to populous group spanning vectors than non-populous group spanning vectors. So the effect of a difference with a large group is shown more strongly than that of a small group in the SMVCIR space.

In practice, we would estimate $\mathbf{K}$ using sample proportions and sample first/second moments. Call this estimate $\mathbf{K}_{\mathbf{n}}$. The standardization would involve sample esti-

mates of the first and second moments as well. We use the familiar $\mathbf{S}$ notation for the sample covariance. We suggest using the scaling transformation on the original data, $\mathbf{z_i} = \text{diag}\left(\mathbf{S}^{-\frac{1}{2}}\right)(\mathbf{x_i} - \bar{\mathbf{x}})$. Alternatively, the Mahalanobis transformation, $\mathbf{z_i} = \mathbf{S}^{-\frac{1}{2}}(\mathbf{x_i} - \bar{\mathbf{x}})$ may be used. Both methods will be discussed in the next chapter.

To go further in our description of the SMVCIR algorithm, we need to use the eigen and singular value matrix decompositions. We briefly review them here.

G. Eigen and Singular Value Decompositions

Let $\mathbf{A}$ be a symmetric $k \times k$ matrix of rank $d \leq k$. The eigen decomposition provides orthonormal k vectors $\mathbf{e_1}, \ldots, \mathbf{e_d}, \ldots, \mathbf{e_k}$ (eigenvectors) and k non-negative scalars $\lambda_1 \geq \ldots \geq \lambda_d > \lambda_{d+1} = \ldots = 0$ (eigenvalues) such that the following holds.

$$\begin{aligned}
\mathbf{E} &= \begin{bmatrix} \mathbf{e_1} & \cdots & \mathbf{e_d} & \cdots & \mathbf{e_k} \end{bmatrix} \\
\mathbf{A} &= \\
&\mathbf{E} \begin{bmatrix} \lambda_1 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \lambda_d & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \mathbf{E}'
\end{aligned} \tag{1.16}$$

Noting that $\mathbf{E}$ is orthogonal since its columns are orthogonal to one another, we

find the following.

$$\mathbf{E}'\mathbf{A}\mathbf{E} = \mathbf{E}'\left(\mathbf{E}\begin{bmatrix} \lambda_1 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \lambda_d & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}\mathbf{E}'\right)\mathbf{E} = \tag{1.17}$$

$$\mathbf{E}'\mathbf{E}\begin{bmatrix} \lambda_1 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \lambda_d & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}\mathbf{E}'\mathbf{E} = \begin{bmatrix} \lambda_1 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \lambda_d & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

We will work with the eigen decomposition of the kernel matrix $\mathbf{K}\mathbf{K}'$ (in theory) and its estimate $\mathbf{K_n}\mathbf{K}'_\mathbf{n}$ (in practice). The eigen decomposition is a continuous function of its input, so since we have a consistent estimate $\mathbf{K_n}$ of $\mathbf{K}$, the eigen decomposition of $\mathbf{K_n}\mathbf{K}'_\mathbf{n}$ provides a consistent estimate of the eigen decomposition of $\mathbf{K}\mathbf{K}'$. There are similar results for the singular value decomposition estimates that we discuss next.

Let $\mathbf{B}$ be a $k \times h$ matrix of rank $d \leq \min(k, h)$. There exist orthonormal k vectors $\mathbf{u_1}, \ldots, \mathbf{u_d}, \ldots, \mathbf{u_k}$ (left singular vectors), orthonormal h vectors $\mathbf{v_1}, \ldots, \mathbf{v_d}, \ldots, \mathbf{v_h}$ (right singular vectors), and $\min(k, h)$ non-negative scalars $\sigma_1 \geq \ldots \geq \sigma_d > \sigma_{d+1} = \ldots = 0$ (singular values) such that the following holds.

$$
\begin{aligned}
\mathbf{U} &= \begin{bmatrix} \mathbf{u_1} & \cdots & \mathbf{u_d} & \cdots & \mathbf{u_k} \end{bmatrix} \\
\mathbf{V} &= \begin{bmatrix} \mathbf{v_1} & \cdots & \mathbf{v_d} & \cdots & \mathbf{v_h} \end{bmatrix} \\
\mathbf{B} &= \\
&\mathbf{U} \begin{bmatrix} \sigma_1 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \sigma_d & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \mathbf{V}'
\end{aligned}
\tag{1.18}
$$

A result similar to (1.17) can be obtained by using the orthogonality of the left and right singular vectors.

$$
\begin{aligned}
&\mathbf{U}'\mathbf{B}\mathbf{V} \\
&= \mathbf{U}'\mathbf{U} \begin{bmatrix} \sigma_1 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \sigma_d & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \mathbf{V}'\mathbf{V} \\
&= \begin{bmatrix} \sigma_1 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \sigma_d & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}
\end{aligned}
\tag{1.19}
$$

Suppose we multiply $\mathbf{B}$ by its transpose.

$$\begin{aligned}
\mathbf{BB}' &= \\
&\mathbf{U}\begin{bmatrix} \sigma_1 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \sigma_d & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}\mathbf{V}'\mathbf{V}\begin{bmatrix} \sigma_1 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \sigma_d & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}\mathbf{U}' \\
&= \mathbf{U}\begin{bmatrix} \sigma_1^2 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \sigma_d^2 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}\mathbf{U}'
\end{aligned} \tag{1.20}$$

$$\begin{aligned}
\mathbf{B}'\mathbf{B} &= \\
&\mathbf{V}\begin{bmatrix} \sigma_1 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \sigma_d & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}\mathbf{U}'\mathbf{U}\begin{bmatrix} \sigma_1 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \sigma_d & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}\mathbf{V}' \\
&= \mathbf{V}\begin{bmatrix} \sigma_1^2 & \cdots & \cdots & \cdots & \cdots & 0 \\ \vdots & \ddots & \cdots & \cdots & \cdots & 0 \\ \vdots & \vdots & \sigma_d^2 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}\mathbf{V}'
\end{aligned} \tag{1.21}$$

So the eigenvalues of the two product matrices, $\mathbf{B}'\mathbf{B}$ and $\mathbf{BB}'$ are the squared singular values of $\mathbf{B}$. The eigenvectors of $\mathbf{B}'\mathbf{B}$ are the right singular vectors of $\mathbf{B}$. The eigenvectors of $\mathbf{BB}'$ are the left-singular vectors of $\mathbf{B}$. This relationship is of particular interest, since our kernel is $\mathbf{KK}'$.

The eigen decomposition provides another useful result. For eigenvector and eigenvalue $i \leq d$ we have the following.

$$\mathbf{A}\mathbf{e_i} = \lambda_i \mathbf{e_i} \tag{1.22}$$

A non-zero matrix $\mathbf{q}$ is in the column space of $\mathbf{A}$ if $\mathbf{At} = \mathbf{q}$ for some vector $\mathbf{t}$. By letting $\mathbf{t} = \lambda_i^{-1}\mathbf{e_i}$, we see that each of the first d eigenvectors are in the column space of $\mathbf{A}$. Since these vectors are orthogonal, they form a basis for the column space of $\mathbf{A}$, which has rank d.

Equation (1.22) holds for higher indices than d, but there the eigenvalue is zero. By the Rank Plus Nullity theorem, the null space of $\mathbf{A}$ has rank $k - d$. Since the eigenvectors of these higher indices are still orthogonal, it follows that they form a basis for the null space of $\mathbf{A}$.

We can apply these results to the product matrix $\mathbf{BB'}$.

$$\mathbf{BB'}\mathbf{u_i} = \sigma_i^2 \mathbf{u_i} \tag{1.23}$$

It follows that each of the first d left singular vector of $\mathbf{B}$ is in the column space of $\mathbf{B}$ (since $\mathbf{B}\left(\mathbf{B'}\mathbf{u_i}\sigma_i^{-2}\right) = \mathbf{u_i}$). And since they are orthogonal, and $\mathbf{B}$ has rank d, they form a basis for the column space of $\mathbf{B}$. Similarly, the last $k - d$ left singular vectors form a basis of the null space of $\mathbf{B}$.

So the eigenvectors of our kernel, $\mathbf{KK'}$, are the left singular vector of spanning set matrix $\mathbf{K}$. They form bases for the column and null spaces of $\mathbf{K}$ as well.

H. Final Translation to SMVCIR Space

The SMVCIR algorithm terminates after translating the standardized predictor coordinates to the SMVCIR spanning set space. There are several different ways that

this could be done. Now that we have learned about the space's decompositions we can discuss each of them.

An orthogonal projection of the standardized predictors onto the SMVCIR space would involve scaling the predictors by a matrix of the form $\mathbf{K}\mathbf{K}^{-}$. After invoking the singular value decomposition, this matrix becomes $\mathbf{U}\mathbf{U}'$. Using this method, we lose some insight on the particular latent dimensions of the column space of $\mathbf{K}$. Information from more than one of the left singular vectors that span the column space is used in making each dimension of the transformed data.

Alternatively, each dimension of the SMVCIR space could be a scalar product (or dot product) of a column from $\mathbf{K}$ and the predictor data vectors. This may lead to heavy dependency between the dimensions of the transformed data, since the columns of $\mathbf{K}$ may be highly dependent. Recall that the columns are centered such that groups of them will add up to zero.

We choose to transform the data by left-multiplying by $\mathbf{U}'$. So each variable of the final transformed data is a dot product of a left singular vector of $\mathbf{K}$ and the standardized predictor data vector. This clearly has smaller dependency problems than the last method, since the left singular vectors are orthogonal to each other. Using the left singular vectors, we isolate the unique differences that span the SMVCIR space. We may view these vectors as sufficient statistics for the information on differences in mean, variance, and covariance that the entire $\mathbf{K}$ matrix has in its columns.

The chosen method of transformation allows inference on the relationship of each unique difference dimension with the data. Each difference is essentially a set of coefficients, representing the location/scale/covariance deviation of a group from the center. The difference coefficients may be obtained via initial investigation of the eigenvectors or later regression of the transformed dimension on the original standardized predictors. By multiplying the predictor data by these coefficients, the trends

that the coefficients represent will be enhanced. Additionally, the weighting ensures that the trend enhancement is fair across the groups, and un-populous but extreme difference trends do not dominate populous and moderate trends.

These left singular vectors are created by forming linear combinations of the vectors in (1.4). Each is formulated so that they are mutually independent. For each of the difference measures: on means, variances, and covariances, the centered parameter vectors will sum to zero. So there is already linear dependency within each measure, and the rank of each set of measure vectors will be at most one less than the number of vectors in the measure.

We can often see the number of linearly independent differences within each measure by visual examination. This is particularly easy when the group proportions are the same. In this case, the unique differences within each measure will show up as subsets of vectors with the same values in that measure. One subset is equivalent to one unique difference. One of the subsets must serve as the base case though. Being what the other measures will add up to, it will not contribute to the number of unique differences.

We can interpret the mean and variance differences easily in this manner. Each unique non-zero subset is a difference in mean/variance from one group to others. We can interpret the vectors of covariance differences in a similar manner, but some duplication occurs. Suppose the first predictor variable differs in covariance from all the others in one group, and that there are no other covariance differences. We find the first column of the covariance difference for that group is unique. But we also find that each of the other columns of that group is unique as well. The differences occur in the first column, and the first row of that group's covariance difference matrix.

It may initially seem like we might be more efficient by only paying attention to the first column of this covariance difference matrix, that the extra dimensions

added by using the first row add redundant information. One possible way to do this is to zero out the upper diagonal of the covariance difference matrices. Then in the last situation we would only obtain a unique difference from the first column of the group's covariance difference matrix.

But suppose rather than the covariance of the first predictor with all others being different in the group we had the covariance of the last predictor with all others being different in the group. Then we would obtain a unique difference from the first through last columns, because we only pay attention to the lower diagonal. So the number of unique covariance difference vectors would vary depending on which predictor was different. We would experience a similar situation if we zeroed out the lower diagonal, with the first and last predictor's role's switched.

So only using the lower or upper diagonal of the covariance difference matrices leads to non-balanced contributions to the SMVCIR space. While it may duplicate information, our method of including a covariance difference vector if it contains either a row/column or column/row index of a covariance difference does ensure that the contribution of individual predictors to the SMVCIR space is balanced.

We can gain some additional insight into the wisdom of our final transformation by appealing to multivariate geometry. An ellipsoid in k dimensions is a set of k dimensional vectors or points $\mathbf{x}$ that satisfy $\mathbf{x}'\mathbf{M}\mathbf{x} = c$ for some non-negative scalar c and symmetric positive semi-definite matrix $\mathbf{M}$.

Muirhead (1982) described how ellipsoids with $\mathbf{M} = \mathrm{Var}[\mathbf{x}]$ were related to the principal components of $\mathbf{X}$, the data matrix with rows corresponding to observations of $\mathbf{x}$. The ith principal component of $\mathbf{X}$ being the ith eigenvector of the matrix $\mathbf{M} = \mathrm{Var}[\mathbf{x}]$ multiplied by the data matrix $\mathbf{X}$. So the principal components are analogous to the transformed variables that we obtain at the end of the SMVCIR algorithm.

According to Muirhead (1982), the principal components "represent a rotation of the coordinate axes to the principal axes of the ellipsoid" for a particular contour. These principal axes are given by the eigenvectors of $\mathbf{M} = \mathrm{Var}[\mathbf{x}]$.

The points satisfying $\mathbf{x}'\mathbf{M}\mathbf{x} = 1$ represent a scaling of the Cartesian coordinates by the variance matrix $\mathbf{M}$. The features found in the points show the features of $\mathbf{M}$, the covariance structure of the underlying data graphically. The principal axes of the ellipsoid are vectors the ellipsoid stretches over. The eigenvalues are the radii of these axes. They measure the quantity of the stretching on the axis. These axes and their radii represent the fundamental feature of the underlying data's covariance structure. They can be interpreted in terms of their elemental values. If the eigenvector with the largest eigenvalue involves only two of the predictor variables, the most prominent features of the covariance structure involves only those two variables, and its direction in those two variables coordinates shows how they are related in the covariance.

So the principal component transformation moves the original data to a space where the features of the covariance structure are more easily seen. The covariance matrix $\mathbf{M} = \mathrm{Var}[\mathbf{x}]$ can be decomposed into the multiplication of two matrices, $\mathbf{M} = (\mathbf{ES})(\mathbf{ES})'$ one the transpose of the other, the $\mathbf{E}$ matrix has the eigenvectors of $\mathbf{M}$ in its columns, while the matrix $\mathbf{S}$ is diagonal with the square root of the eigenvalues of $\mathbf{M}$ on its diagonal. Clearly by (1.16), $\mathbf{M} = \mathbf{ESS}'\mathbf{E}' = \mathbf{EDE}'$, where $\mathbf{D}$ is diagonal and contains the eigenvalues of $\mathbf{M}$ along its diagonal.

So the features of $\mathbf{M}$ are functions of the components of the matrix $\mathbf{ES}$. The matrix $\mathbf{E}$ must be $k \times k$ and we earlier assumed that $\mathbf{S}$ was $k \times k$ and diagonal. By adding zero columns to the right, we could redefine $\mathbf{S}$ to be $k \times h$, where h is at least equal to the rank of $\mathbf{M}$. Under this new definition no new information would be introduced to $\mathbf{M}$ and none would be lost.

We could also post multiply $\mathbf{ES}$ by the transpose of an orthogonal matrix $h \times h$,

$\mathbf{V}$. We find that $\mathbf{M} = \mathbf{ESV'VS'E'} = \mathbf{EDE'}$. So the post-multiplication by $\mathbf{V'}$ neither adds nor removes information from $\mathbf{M}$. Recall our singular value decomposition work. This tells us that M is completely determined by the left singular vectors and singular values of the matrix $\mathbf{T} = \mathbf{ESV'}$. Furthermore, $\mathbf{x'Mx} = \mathbf{x'TT'x} = (\mathbf{T'x})' (\mathbf{T'x})$.

So the central features of the $\mathbf{M}$ matrix are functions of the column space of the matrix $\mathbf{T} = \mathbf{ESV'}$ and its singular values. So by rotating the original data to the principal axes of the $\mathbf{x'Mx} = 1$ ellipsoid through the principal components, we are rotating the data to the central features of the column space of $\mathbf{T}$, and the strength of those features is reflected by the singular values of $\mathbf{T}$, which are the square roots of the eigenvalues of $\mathbf{M}$.

In the case of SMVCIR, we obtain the same interpretation by changing $\mathbf{T} := \mathbf{K}$ and thus $\mathbf{M} = \mathbf{KK'}$. So by rotating the standardized original data to the principal axes of the $(\mathbf{K'z})' (\mathbf{K'z}) = 1$ ellipsoid through the principal components, we are rotating the data to the central features of the column space of $\mathbf{K}$, and the strength of those features is reflected by the singular values of $\mathbf{K}$. This also means that the most discriminating dimensions will be those corresponding to the highest eigenvalues and sorted toward the beginning.

I. Original SMVCIR Ordering Procedure

Now we have described the basic SMVCIR algorithm and explained how to interpret its results. A very useful augmentation is developed in the Sheather et al. (2008) paper. This augmentation replaces $\mathbf{K}$ with a k row sub-matrix composed of r columns from $\mathbf{K}$.

In section 12.2 of Golub and Loan (1996), details are given on how to approximate a matrix of column vectors using subsets of these column vectors. The target of the

approximation is the distance between an arbitrary vector and an element of the column space of the approximated matrix. They seek to find the best vector to right multiply the approximated matrix by to minimize this distance.

Given a vector $\mathbf{b}$, finding a vector $\mathbf{x}$ to minimize the residual $|\mathbf{Ax} - \mathbf{b}|$ is trivial. What Golub and Loan (1996) wished to find was an r-column approximation $\mathbf{A_r}$ to $\mathbf{A}$ such that the residual found using $\mathbf{A_r}$ is close to that found using $\mathbf{A}$. In calculating these residuals, we are essentially calculating the distance from $\mathbf{b}$ to the column spaces of $\mathbf{A_r}$ or $\mathbf{A}$. So by finding $\mathbf{A_r}$ to minimize the difference between the residuals, we are finding $\mathbf{A_r}$ closest to $\mathbf{A}$ in column space.

Golub and Loan (1996) advised that r be chosen as the rank estimate of $\mathbf{A}$. This is the order of smallest non-zero eigenvalue or singular value of $\mathbf{A}$. One possibility for approximating $\mathbf{A}$ was by using the first r left and right singular vectors. Recalling our notation from (1.18) we obtain the following.

$$\begin{aligned} \mathbf{U_r} &= \begin{bmatrix} \mathbf{u_1} & \cdots & \mathbf{u_r} \end{bmatrix} \\ \mathbf{V_r} &= \begin{bmatrix} \mathbf{v_1} & \cdots & \mathbf{v_r} \end{bmatrix} \\ \mathbf{A_r} &= \\ & \mathbf{U_r} \begin{bmatrix} \sigma_1 & \cdots & \cdots \\ \vdots & \ddots & \cdots \\ \vdots & \vdots & \sigma_r \end{bmatrix} \mathbf{V_r'} = \textstyle\sum_{i=1}^{r} \sigma_i \mathbf{u_i} \mathbf{v_i'} \end{aligned} \tag{1.24}$$

There is another result documented in Section 1.1.16 of Golub and Loan (1996) that sheds intuitive light on this approximation.

$$\begin{aligned} \mathbf{A} &= \begin{bmatrix} \mathbf{a_1} & \cdots & \mathbf{a_p} \end{bmatrix} \\ \mathbf{B} &= \begin{bmatrix} \mathbf{b_1} & \cdots & \mathbf{b_p} \end{bmatrix} \\ \mathbf{AB'} &= \textstyle\sum_{k=1}^{p} \mathbf{a_i} \mathbf{b_i'} \end{aligned} \tag{1.25}$$

So a matrix multiplication can be written as the sum of the matrix products of the component vectors in each matrix. The singular value decomposition is a weighted (by the singular values) sum of the matrix products of the left and right singular vectors. The most important components are weighted the highest. So the approximation only considers the r highest weighted components.

Recall that in practice, $\mathbf{A}$ would be an estimate of a matrix, like our $\mathbf{K_n}$ is an estimate of $\mathbf{K}$. A method for choosing r in this situation is given in Sheather et al. (2008). A threshold value is chosen based on the percentage of the total singular value sum that an individual singular value comprises. When this percentage is too small for a singular value i and too large or just right for the previously ordered singular value $i-1$, then r is chosen to be $i-1$.

We can be very flexible with our threshold. We may even ignore small singular values that we assume match with true non-zero singular values in the population matrix $\mathbf{A}$. These may correspond to randomly noisy dimensions of the column space of $\mathbf{A}$.

So this first method approximates $\mathbf{A}$ by using its most "loud" singular components. An alternative method does not use the singular vectors of $\mathbf{A}$ in its final approximation calculation. This method, which is employed by Sheather et al. (2008) orders the columns of $\mathbf{A}$ in terms of their linear independence and then attains $\mathbf{A_r}$ by keeping the first r columns. In this method we ignore the "redundant" column space dimensions of $\mathbf{A}$, rather than the randomly noisy dimensions.

Both approaches can give similar and useful results. Bounds on the difference between the residuals found under both methods can be found section 12.2 of Golub and Loan (1996).

In the inference on the number of dimensions in the SMVCIR space, we focus on statistical significance instead of practical significance. We will not use the approx-

imation method as we develop our dimensionality tests. It is certainly useful, and the number of dimensions it yields may often coincide with the results of our dimensionality tests. In practice it may be useful to perform both algorithms in concert and make a decision on the number of dimensions of the SMVCIR space based on both sets of results. This possibility should be investigated in further study. We will return to this notion in the last chapter as we discuss future research directions.

We end the chapter by summarizing the SMVCIR algorithm pictorially in Figure 4. It was simpler to describe the centering of the group means and group variance/covariances together in the chapter. Matrix multiplication distributes over matrix addition/subtraction, so we may center the group variance/covariances before or after standardization of the parameters.

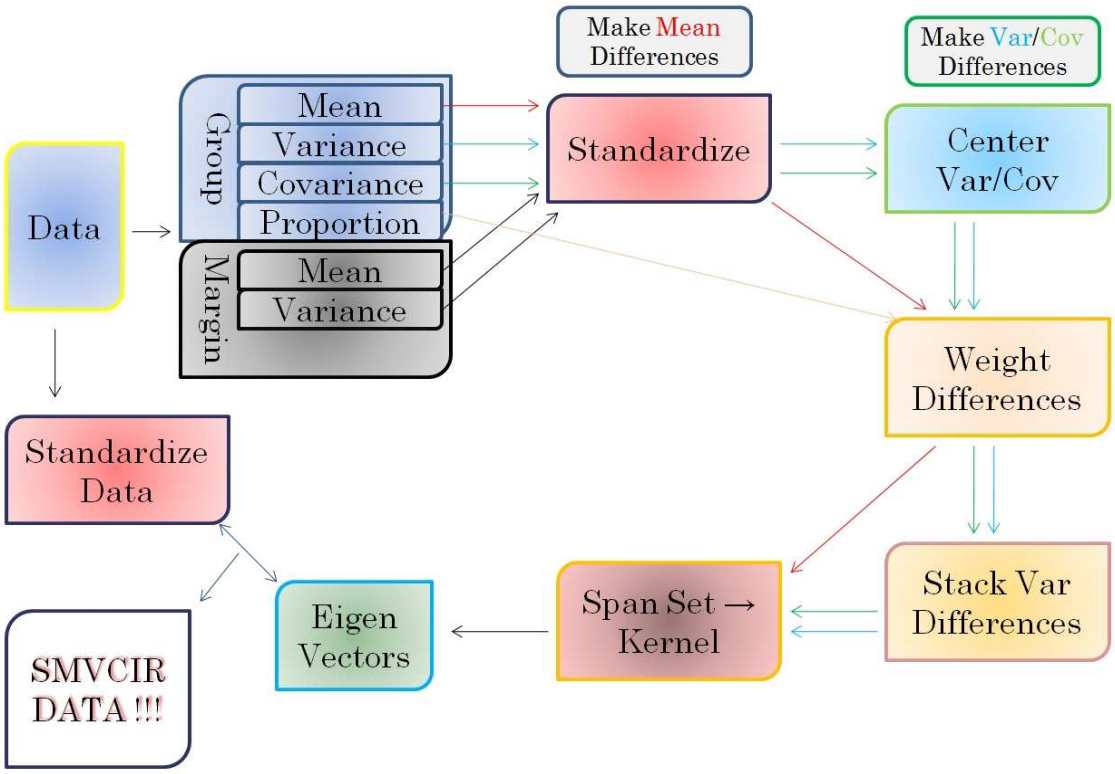

Figure 4. SMVCIR Algorithm

CHAPTER II

SMVCIR STANDARDIZATION METHODS

We standardize the data to solve two problems, marginal scaling and marginal linear dependency. When there is marginal linear dependency among the variables, they are correlated in $\mathbf{\Sigma}$ and they may remain correlated under the translation to SMVCIR coordinates, so visualization and interpretation of the new coordinates may be difficult. When the marginal scaling drastically differs in $\mathbf{\Sigma}$ for different predictors, their transformed scales may drastically differ as well. Visualization may be easier when coordinates share similar scales.

Using the Mahalanobis transformation, $\mathbf{z_i} = \mathbf{S}^{-\frac{1}{2}}(\mathbf{x_i} - \bar{\mathbf{x}})$ we fix both problems and also center the marginal mean to zero. But this transformation can introduce differences between the groups that were not found in the original data.

Alternatively, we could use the scaling transformation. Under this transformation, we replace the observations $\mathbf{x_i}$ with $\mathbf{z_i} = \text{diag}\left(\mathbf{S}^{-\frac{1}{2}}\right)(\mathbf{x_i} - \bar{\mathbf{x}})$. Both these transformations are explained fully in Mardia et al. (1979). This transformation will solve the marginal scaling problem but will not solve any marginal linear dependency problems. It will also not introduce new differences in location/scale/covariance for SMVCIR to detect.

We begin this chapter with a brief discussion of how standardization is handled in the SAVE and SIR algorithms. Then we show how it works in SMVCIR. We will show the Mahalanobis transformation can introduce new group differences via several examples. We also will show that the scaling transformation does not introduce new group differences empirically through the same examples and formally via proof. We introduce the notion of strong conversion of the mean, variance, covariance parameters for this purpose. At the end of the chapter we will show how

the Mahalanobis transformation can solve both marginal scaling and marginal linear dependency problems, and how the scaling transformation solves marginal scaling problems. The use of the Mahalanobis transformation is finally advocated in the case the the scaling transformation leads to highly correlated SMVCIR coordinates. If this does not occur, then the scaling transformation is advocated.

A. SAVE and SIR Standardization

The Mahalanobis transformation commits no offenses when used in the SAVE and SIR methods. SIR assumes equal variance across the groups (Li (1991)). Both SIR and SAVE target the central subspace of the response conditioned on the predictors. As proven in proposition 6.1 of (Cook (1998)), the spanning set of this central subspace is obtained from the spanning set of the standardized predictor central subspace via a multiplication by $\mathbf{S}^{-\frac{1}{2}}$. So the two spaces are linked by the invertible $\mathbf{S}^{-\frac{1}{2}}$ multiplication transformation. It is not a concern if differences are introduced via the standardization then, because they are removed when moving back to the original predictor scale for graphing of the data along the SAVE or SIR dimensions.

SMVCIR does not target the central subspace. Its stacking of variance parameters in vectors while leaving the covariance matrices with zero diagonal entries is unconventional. It does not estimate the space spanned by mean differences, and variance-covariance matrix column differences of SAVE, $\mathbf{\Omega_{SAVE}}$ (defined in Sheather et al. (2008)). This suggests that the simple $\mathbf{S}^{-\frac{1}{2}}$ link between the standardized and original predictor SMVCIR spanning sets may not exist. We will show that the equally simple diag $\left(\mathbf{S}^{-\frac{1}{2}}\right)$ link does exist and is useful.

B. Mahalanobis Problem and Strong Conversion

We will first explain how the Mahalanobis transformation can introduce new differences in location/scale/covariance between the groups. This essentially generates noise dimensions in the SMVCIR space. They may not be influential, but they will be there, and the tests of dimensionality that we develop may detect these dimensions and lead us to infer that there are more "real" (in original data space) dimensions than we should.

The SMVCIR algorithm works on the centered group parameter vectors and matrices, detecting deviations from the center through these group difference vectors and matrices. The invariance in the ranking relationships among the group means, variances, and covariances should be maintained when we move to a standardized scale for the predictors.

This means, for groups i and l, predictors x_j and x_m, and operator $\sim$ from $=, >, <$ the following hold.

$$\begin{aligned} \mu^c_{ij} &\sim \mu^c_{lj} \Leftrightarrow \\ \mu_{ij} &\sim \mu_{lj} \Leftrightarrow \\ \mu^z_{ij} &\sim \mu^z_{lj} \Leftrightarrow \\ \mu^{zc}_{ij} &\sim \mu^{zc}_{lj} \end{aligned} \tag{2.1}$$

$$\begin{aligned} \sigma^c_{ijm} &\sim \sigma^c_{ljm} \Leftrightarrow \\ \sigma_{ijm} &\sim \sigma_{ljm} \Leftrightarrow \\ \sigma^z_{ijm} &\sim \sigma^z_{ljm} \Leftrightarrow \\ \sigma^{zc}_{ijm} &\sim \sigma^{zc}_{ljm} \end{aligned} \tag{2.2}$$

The first and third dual implications in (2.1) and (2.2) are obviously true, since the centered parameters are only location shifted. We have to show that our method of

standardization ensures that the second dual implication is true. The rank/difference relationships must be maintained when we switch scales.

Strong Conversion Principle

Is satisfied by Affine transformation $(\mathbf{A}_{k\times k},\mathbf{b}_k)$ on
Data $\mathbf{x}$ from populations $(\boldsymbol{\mu_1}, \boldsymbol{\Sigma_1}), \ldots (\boldsymbol{\mu_g}, \boldsymbol{\Sigma_g})$ (2.3)
if it maintains (2.1) and (2.2) in
$\mathbf{x} \to \mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{b}$

The coordinate transformation used in SMVCIR should satisfy (2.3). Since it represents a global location shift, the choice of $\mathbf{b}$ is not relevant in determining whether a transformation satisfies the Strong Conversion Principle. We will now investigate how to choose an $\mathbf{A}$ matrix to satisfy the principle.

First we focus on the group means.

$$\begin{aligned}
&\mathrm{E}\left[\mathbf{A}\mathbf{x} + \mathbf{b} | y = i\right] = \\
&\mathbf{A}\mathrm{E}\left[\mathbf{x} | y = i\right] + \mathbf{b} = \\
&\mathbf{A}\boldsymbol{\mu_i} + \mathbf{b} = \\
&\begin{bmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{bmatrix} \begin{bmatrix} \mu_{i1} \\ \vdots \\ \mu_{ik} \end{bmatrix} + \begin{bmatrix} b_1 \\ \vdots \\ b_k \end{bmatrix} = \\
&\begin{bmatrix} a_{11}\mu_{i1} + \ldots + a_{1k}\mu_{ik} + b_1 \\ \vdots \\ a_{k1}\mu_{i1} + \ldots + a_{kk}\mu_{ik} + b_k \end{bmatrix}
\end{aligned} \tag{2.4}$$

So after translation, each component of the standardized group mean depends on all of the components of the un-standardized group mean via the coefficients in a row of $\mathbf{A}$. So to ensure strong conversion for two group mean components, we have to examine a linear combination of each of the group mean components with coefficients

from the **A** matrix.

We must insure that $\mu_{ij} \sim \mu_{lj}$ iff $a_{j1}\mu_{i1} + \ldots + a_{jk}\mu_{ik} + b_j \sim a_{j1}\mu_{l1} + \ldots + a_{jk}\mu_{lk} + b_j$. The irrelevance of the **b** vector is clear here, since it will cancel on both sides. The complication increases when we consider strong conversion for the variance parameters.

$\text{Var}\left[\mathbf{Ax} + \mathbf{b} | y = i\right] =$

$\mathbf{A}\text{Var}\left[\mathbf{x} | y = i\right]\mathbf{A}' =$

$\mathbf{A\Sigma_i A}' =$

$$\begin{bmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{bmatrix} \begin{bmatrix} \sigma_{i11} & \cdots & \sigma_{i1k} \\ \vdots & \ddots & \vdots \\ \sigma_{ik1} & \cdots & \sigma_{ikk} \end{bmatrix} \begin{bmatrix} a_{11} & \cdots & a_{k1} \\ \vdots & \ddots & \vdots \\ a_{1k} & \cdots & a_{kk} \end{bmatrix} =$$

$$\begin{bmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{bmatrix} \begin{bmatrix} a_{11}\sigma_{i11} + \ldots + a_{1k}\sigma_{i1k} & \cdots & a_{k1}\sigma_{i11} + \ldots + a_{kk}\sigma_{i1k} \\ \vdots & \ddots & \vdots \\ a_{11}\sigma_{ik1} + \ldots + a_{1k}\sigma_{ikk} & \cdots & a_{k1}\sigma_{ik1} + \ldots + a_{kk}\sigma_{ikk} \end{bmatrix} =$$

$$\begin{bmatrix} \begin{bmatrix} a_{11}\left(a_{11}\sigma_{i11} + \ldots + a_{1k}\sigma_{i1k}\right) + a_{1k}\left(a_{11}\sigma_{ik1} + \ldots + a_{1k}\sigma_{ikk}\right) \\ \vdots \\ a_{11}\left(a_{k1}\sigma_{i11} + \ldots + a_{kk}\sigma_{i1k}\right) + a_{1k}\left(a_{k1}\sigma_{ik1} + \ldots + a_{kk}\sigma_{ikk}\right) \end{bmatrix}' \\ \vdots \\ \begin{bmatrix} a_{k1}\left(a_{11}\sigma_{i11} + \ldots + a_{1k}\sigma_{i1k}\right) + a_{kk}\left(a_{11}\sigma_{ik1} + \ldots + a_{1k}\sigma_{ikk}\right) \\ \vdots \\ a_{k1}\left(a_{k1}\sigma_{i11} + \ldots + a_{kk}\sigma_{i1k}\right) + a_{kk}\left(a_{k1}\sigma_{ik1} + \ldots + a_{kk}\sigma_{ikk}\right) \end{bmatrix}' \end{bmatrix} \tag{2.5}$$

So after translation, each component of the standardized group variance depends on all of the components of the un-standardized group variance. Ensuring strong conversion for the variance parameters has the potential to be very complicated.

The complication comes from the dependence of the individual standardized parameters on multiple of the un-standardized parameters. This dependence is implemented in the structure of $\mathbf{A}$. So by simplifying the structure of $\mathbf{A}$, we reduce the complexity of the dependence and strong conversion is more feasible and simple to show.

If we make $\mathbf{A}$ diagonal and positive definite, we allow for powerful scaling of the means and variance parameters, and ensure strong conversion. We will prove that $\mathbf{A}$ diagonal and positive definite leads to strong conversion, and then show examples where a non-diagonal $\mathbf{A}$ does not yield strong conversion. The square root of the inverse of the marginal variance is the $\mathbf{A}$ of choice for SIR and SAVE. In the examples that follow the proof, we will demonstrate how this selection does not yield strong conversion. We will also show how using the square root of the inverse of the marginal variance matrix's diagonal, which we suggest for SMVCIR, does yield strong conversion.

We begin our proof by showing what the mean of the converted variables is when $\mathbf{A}$ is diagonal and positive definite.

$$
\begin{aligned}
&\mathrm{E}\left[\mathbf{A}\mathbf{x}+\mathbf{b}|y=i\right]= \\
&\mathbf{A}\mathrm{E}\left[\mathbf{x}|y=i\right]+\mathbf{b}= \\
&\mathbf{A}\boldsymbol{\mu}_{\mathbf{i}}+\mathbf{b}= \\
&\begin{bmatrix} a_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_{kk} \end{bmatrix} \begin{bmatrix} \mu_{i1} \\ \vdots \\ \mu_{ik} \end{bmatrix} + \begin{bmatrix} b_1 \\ \vdots \\ b_k \end{bmatrix} = \\
&\begin{bmatrix} a_{11}\mu_{i1}+b_1 \\ \vdots \\ a_{kk}\mu_{ik}+b_k \end{bmatrix}
\end{aligned}
\tag{2.6}
$$

For strong conversion on the mean parameters, it must be the case that $\mu_{ij} \sim \mu_{lj}$ if and only if $a_{jj}\mu_{ij} + b_j \sim a_{jj}\mu_{lj} + b_j$ for each predictor index j and groups indices i and l. Since the operator $\sim$ is from $=$, $>$, or $<$, and $a_{jj} > 0$, this is true. So using a diagonal and positive definite $\mathbf{A}$ ensures strong conversion for the mean parameters.

$$
\begin{aligned}
&\mathrm{Var}\left[\mathbf{Ax} + \mathbf{b} | y = i\right] = \\
&\mathbf{A}\mathrm{Var}\left[\mathbf{x} | y = i\right]\mathbf{A}' = \\
&\mathbf{A\Sigma_i A}' = \\
&\begin{bmatrix} a_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_{kk} \end{bmatrix}\begin{bmatrix} \sigma_{i11} & \cdots & \sigma_{i1k} \\ \vdots & \ddots & \vdots \\ \sigma_{ik1} & \cdots & \sigma_{ikk} \end{bmatrix}\begin{bmatrix} a_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_{kk} \end{bmatrix} = \\
&\begin{bmatrix} a_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & a_{kk} \end{bmatrix}\begin{bmatrix} a_{11}\sigma_{i11} & \cdots & a_{kk}\sigma_{i1k} \\ \vdots & \ddots & \vdots \\ a_{11}\sigma_{ik1} & \cdots & a_{kk}\sigma_{ikk} \end{bmatrix} = \\
&\begin{bmatrix} a_{11}^2\sigma_{i11} & \cdots & a_{11}a_{kk}\sigma_{i1k} \\ \vdots & \ddots & \vdots \\ a_{kk}a_{11}\sigma_{ik1} & \cdots & a_{kk}^2\sigma_{ikk} \end{bmatrix}
\end{aligned}
\tag{2.7}
$$

For strong conversion on the variance parameters, it must the case that $\sigma_{ijm} \sim \sigma_{ljm}$ if and only if $a_{jj}a_{mm}\sigma_{ijm} \sim a_{jj}a_{mm}\sigma_{ljm}$ for predictor indices j and m and group indices i and l. Since the operator $\sim$ is from $=$, $>$, and $<$, and $a_{jj}, a_{mm} > 0$, this is true. So strong conversion for the variance parameters is insured. We conclude that the following theorem is true.

Strong Conversion Theorem

Affine transformation $(\mathbf{A}_{k\times k}, \mathbf{b}_k)$ satisfies the Strong Conversion Principle if $\mathbf{A}$ is diagonal and positive definite. (2.8)

There are several potentially good choices for $(\mathbf{A}, \mathbf{b})$. We might choose for $\mathbf{A}$ to have the inverted pooled standard deviation along its diagonal, and then pre-center the new group means using $\mathbf{b} = -\mathbf{A}\boldsymbol{\mu}$. We could also choose $\mathbf{A}$ to have the inverted marginal standard deviations along its diagonal and then pre-center using $\mathbf{b} = -\mathbf{A}\boldsymbol{\mu}$ as well. Let us test both these standardizations using the population data in (1.10).

C. Standardization Example 1, from Equation 1.10

We start with the pooled standard deviation option.

$$\begin{aligned} &\mathrm{E}\left[\mathbf{A}_{pooled}\mathbf{x} + \mathbf{b}_{pooled} | y = 1\right] = \\ &-\begin{bmatrix} 1.1^{-.5} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 100 \\ 100 \\ 75 \end{bmatrix} = \begin{bmatrix} -95.34626 \\ -100 \\ -75 \end{bmatrix} \end{aligned} \tag{2.9}$$

$$\begin{aligned} &\mathrm{E}\left[\mathbf{A}_{\text{pooled}}\mathbf{x} + \mathbf{b}_{pooled} | y = 2\right] = \\ &\begin{bmatrix} 1.1^{-.5} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 100 \\ 100 \\ 75 \end{bmatrix} = \begin{bmatrix} 95.34626 \\ 100 \\ 75 \end{bmatrix} \end{aligned} \tag{2.10}$$

$$\begin{aligned} &\mathrm{Var}\left[\mathbf{A}_{pooled}\mathbf{x} + \mathbf{b}_{pooled} | y = 1\right] = \\ &\begin{bmatrix} 1.1^{-.5} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1.1^{-.5} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \\ &\begin{bmatrix} .\bar{9}\bar{0} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \end{aligned} \tag{2.11}$$

$$\begin{aligned}&\text{Var}\left[\mathbf{A}_{pooled}\mathbf{x}+\mathbf{b}_{pooled}|y=2\right]=\\&\begin{bmatrix}1.1^{-.5}&0&0\\0&1&0\\0&0&1\end{bmatrix}\begin{bmatrix}1.2&0&0\\0&1&0\\0&0&1\end{bmatrix}\begin{bmatrix}1.1^{-.5}&0&0\\0&1&0\\0&0&1\end{bmatrix}=\\&\begin{bmatrix}1.\bar{0}\bar{9}&0&0\\0&1&0\\0&0&1\end{bmatrix}\end{aligned}\tag{2.12}$$

When we center the new standardized parameters, the means remain the same, and the variances become the following.

$$\mathbf{\Sigma}_{\mathbf{1}}^{zc}=\begin{bmatrix}-.\bar{0}\bar{9}&0&0\\0&0&0\\0&0&0\end{bmatrix}\tag{2.13}$$

$$\mathbf{\Sigma}_{2}^{zc}=\begin{bmatrix}.\bar{0}\bar{9}&0&0\\0&0&0\\0&0&0\end{bmatrix}\tag{2.14}$$

The wide difference between the variance and mean differences (centered variances and means) is very apparent here. It may be awkward to interpret the mean differences as being standardized here, because of their large magnitude. Using the marginal standard deviations in the other standardization method will solve this problem.

$$\begin{aligned}&\text{E}\left[\mathbf{A}_{\text{marginal}}\mathbf{x}+\mathbf{b}_{\text{marginal}}|y=1\right]=\\&-\begin{bmatrix}10001.1^{-.5}&0&0\\0&10001^{-.5}&0\\0&0&5626^{-.5}\end{bmatrix}\begin{bmatrix}100\\100\\75\end{bmatrix}\approx\begin{bmatrix}-.99995\\-.99995\\-.99991\end{bmatrix}\end{aligned}\tag{2.15}$$

$$\begin{aligned}&\text{E}\left[\mathbf{A}_{\text{marginal}}\mathbf{x}+\mathbf{b}_{\text{marginal}}|y=2\right]=\\&\begin{bmatrix}10001.1^{-.5}&0&0\\0&10001^{-.5}&0\\0&0&5626^{-.5}\end{bmatrix}\begin{bmatrix}100\\100\\75\end{bmatrix}\approx\begin{bmatrix}.99995\\.99995\\.99991\end{bmatrix}\end{aligned}\tag{2.16}$$

$$\begin{aligned}
&\mathrm{Var}\left[\mathbf{A}_{marginal}\mathbf{x}+\mathbf{b}_{marginal}|y=1\right]= \\
&\begin{bmatrix} 10001.1^{-.5} & 0 & 0 \\ 0 & 10001^{-.5} & 0 \\ 0 & 0 & 5626^{-.5} \end{bmatrix}\begin{bmatrix} 10001.1^{-.5} & 0 & 0 \\ 0 & 10001^{-.5} & 0 \\ 0 & 0 & 5626^{-.5} \end{bmatrix}= \\
&\begin{bmatrix} 10001.1^{-1} & 0 & 0 \\ 0 & 10001^{-1} & 0 \\ 0 & 0 & 5626^{-1} \end{bmatrix}
\end{aligned} \tag{2.17}$$

$$\begin{aligned}
&\mathrm{Var}\left[\mathbf{A}_{marginal}\mathbf{x}+\mathbf{b}_{marginal}|y=2\right]= \\
&\begin{bmatrix} 10001.1^{-.5} & 0 & 0 \\ 0 & 10001^{-.5} & 0 \\ 0 & 0 & 5626^{-.5} \end{bmatrix}\begin{bmatrix} 1.2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \\
&\begin{bmatrix} 10001.1^{-.5} & 0 & 0 \\ 0 & 10001^{-.5} & 0 \\ 0 & 0 & 5626^{-.5} \end{bmatrix} \\
&= \\
&\begin{bmatrix} 1.2/10001.1 & 0 & 0 \\ 0 & 10001^{-1} & 0 \\ 0 & 0 & 5626^{-1} \end{bmatrix}
\end{aligned} \tag{2.18}$$

Now the mean parameters have more "standardized" looking values, and the variance differences will be incredibly small instead of showing deviation around .1.

In this example and in general, the pooled standard deviations will always be less than the marginal standard deviations. Recall the partitioning of the marginal variance. The pooled standard deviations are the diagonal elements of (1.7). The marginal standard deviations are those diagonal elements added to the diagonal elements of (1.8), all of which are non-negative. So to guarantee that the scale of our standardized variables is suitably small, we suggest standardizing by choosing $\mathbf{A}$ to have the inverted marginal standard deviations along its diagonal and pre-centering with $\mathbf{b}=-\mathbf{A}\boldsymbol{\mu}$.

In our current example, let us see what happens when we standardize using the square-root inverse of the marginal variance. We calculate the square-root of the

inverse of $\mathbf{\Sigma}$ using an eigen decomposition.

$$\mathbf{\Sigma}^{-\frac{1}{2}} \approx \begin{bmatrix} +.5944111 & -.3764266 & -.2823200 \\ -.3764266 & +.6049119 & -.2963160 \\ -.2823200 & -.2963160 & +.7777630 \end{bmatrix} \tag{2.19}$$

Now we will compute the standardized parameters.

$$\begin{aligned} &\boldsymbol{\mu}_{\mathbf{1}}^{\mathbf{z}} = \mathbf{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}_{\mathbf{1}} \approx \\ &- \begin{bmatrix} +.5944111 & -.3764266 & -.2823200 \\ -.3764266 & +.6049119 & -.2963160 \\ -.2823200 & -.2963160 & +.7777630 \end{bmatrix} \begin{bmatrix} 100 \\ 100 \\ 75 \end{bmatrix} \approx \\ &\begin{bmatrix} -.624450 \\ -.624830 \\ -.468625 \end{bmatrix} \end{aligned} \tag{2.20}$$

$$\begin{aligned} &\boldsymbol{\mu}_{\mathbf{2}}^{\mathbf{z}} = \mathbf{\Sigma}^{-\frac{1}{2}}\boldsymbol{\mu}_{\mathbf{2}} \approx \\ &\begin{bmatrix} +.5944111 & -.3764266 & -.2823200 \\ -.3764266 & +.6049119 & -.2963160 \\ -.2823200 & -.2963160 & +.7777630 \end{bmatrix} \begin{bmatrix} 100 \\ 100 \\ 75 \end{bmatrix} \approx \\ &\begin{bmatrix} .624450 \\ .624830 \\ .468625 \end{bmatrix} \end{aligned} \tag{2.21}$$

$$\begin{aligned} &\mathbf{\Sigma}_{\mathbf{1}}^{\mathbf{z}} = \mathbf{\Sigma}^{-\frac{1}{2}}\mathbf{\Sigma}_{\mathbf{1}}\mathbf{\Sigma}^{-\frac{1}{2}} \approx \\ &\begin{bmatrix} +.5944111 & -.3764266 & -.2823200 \\ -.3764266 & +.6049119 & -.2963160 \\ -.2823200 & -.2963160 & +.7777630 \end{bmatrix}^2 = \\ &\begin{bmatrix} +.5747261 & -.3678011 & -.2758510 \\ -.3678011 & +.5954186 & -.3034359 \\ -.2758510 & -.3034359 & +.7724230 \end{bmatrix} \end{aligned} \tag{2.22}$$

$$\begin{aligned}
&\boldsymbol{\Sigma}^{-\frac{1}{2}}\boldsymbol{\Sigma_2}\boldsymbol{\Sigma}^{-\frac{1}{2}} \approx \\
&\begin{bmatrix} +.5944111 & -.3764266 & -.2823200 \\ -.3764266 & +.6049119 & -.2963160 \\ -.2823200 & -.2963160 & +.7777630 \end{bmatrix} \begin{bmatrix} +.7132930 & -.3764266 & -.2823200 \\ -.4517120 & +.6049119 & -.2963160 \\ -.3387840 & -.2963160 & +.7777630 \end{bmatrix} \\
&\approx \\
&\begin{bmatrix} +.6453910 & -.4125516 & -.3094138 \\ -.4125516 & +.6237580 & -.2821814 \\ -.3094138 & -.2821814 & +.7883640 \end{bmatrix}
\end{aligned} \tag{2.23}$$

It is immediately clear that strong conversion is not maintained for nearly all of the variance parameters. The group covariances, originally zero and equal across groups one and two, have been changed to non-zero values and are no longer equal across the two groups. Between group differences in the second and third component of the variances have been introduced as well.

SMVCIR stacks all the variance parameters together in its spanning set, so these new variance differences are not a great cause for concern (since the first component was already different). However, three covariance difference vectors will be added to the spanning set because of the new covariance component differences.

We will compute the centered variance matrices and then show the spanning set for SMVCIR. The spanning set is represented by a matrix whose columns are the dimensions of the SMVCIR space, the spanning set vectors.

$$\begin{aligned}
&\sum_{i=1}^{g} p_i \boldsymbol{\Sigma}_{\mathbf{i}}^{\mathbf{z}} \approx \\
&\frac{1}{2}\begin{bmatrix} +.5747261 & -.3678011 & -.2758510 \\ -.3678011 & +.5954186 & -.3034359 \\ -.2758510 & -.3034359 & +.7724230 \end{bmatrix} + \\
&\frac{1}{2}\begin{bmatrix} +.6453910 & -.4125516 & -.3094138 \\ -.4125516 & +.6237580 & -.2821814 \\ -.3094138 & -.2821814 & +.7883640 \end{bmatrix} \\
&\approx \begin{bmatrix} +.6100586 & -.3901764 & -.2926324 \\ -.3901764 & +.6095883 & -.2928087 \\ -.2926324 & -.2928087 & +.7803935 \end{bmatrix}
\end{aligned} \tag{2.24}$$

$$\begin{aligned}
&\boldsymbol{\Sigma}_{\mathbf{1}}^{\mathbf{zc}} \approx \\
&\left[\begin{array}{ccc} +.5747261 & -.3678011 & -.2758510 \\ -.3678011 & +.5954186 & -.3034359 \\ -.2758510 & -.3034359 & +.7724230 \end{array}\right] - \\
&\left[\begin{array}{ccc} +.6100586 & -.3901764 & -.2926324 \\ -.3901764 & +.6095883 & -.2928087 \\ -.2926324 & -.2928087 & +.7803935 \end{array}\right] \\
&\approx \left[\begin{array}{ccc} -.0353325 & +.0223752 & +.0167814 \\ +.0223752 & -.0141697 & -.0106273 \\ +.0167814 & -.0106273 & -.0079705 \end{array}\right]
\end{aligned} \tag{2.25}$$

$$\begin{aligned}
&\boldsymbol{\Sigma}_{\mathbf{2}}^{\mathbf{zc}} \approx \\
&\left[\begin{array}{ccc} +.6453910 & -.4125516 & -.3094138 \\ -.4125516 & +.6237580 & -.2821814 \\ -.3094138 & -.2821814 & +.7883640 \end{array}\right] - \\
&\left[\begin{array}{ccc} +.6100586 & -.3901764 & -.2926324 \\ -.3901764 & +.6095883 & -.2928087 \\ -.2926324 & -.2928087 & +.7803935 \end{array}\right] \\
&\approx \left[\begin{array}{ccc} +.0353325 & -.0223752 & -.0167814 \\ -.0223752 & +.0141697 & +.0106273 \\ -.0167814 & +.0106273 & +.0079705 \end{array}\right]
\end{aligned} \tag{2.26}$$

Given this, the SMVCIR spanning set (ignoring the group weights, since they are equal) is given by the following.

$$\mathbf{Sp}' \approx \left[\begin{array}{ccc} -.624450 & -.624830 & -.468625 \\ +.624450 & +.624830 & +.468625 \\ -.035333 & -.014170 & -.007971 \\ +.035333 & -.014170 & +.007971 \\ 0 & +.022375 & +.016781 \\ +.022375 & 0 & -.010627 \\ +.016781 & -.010627 & 0 \\ 0 & -.022375 & -.016781 \\ -.022375 & 0 & +.010627 \\ -.016781 & .0106273 & 0 \end{array}\right]' \tag{2.27}$$

Here we stacked the variances together after the means. Treating the first group

as a baseline, we obtain an equivalent spanning set by dropping that group's vectors.

$$\mathbf{Sp}' \approx \begin{bmatrix} +.624450 & +.624830 & +.468625 \\ +.035333 & +.014170 & +.007971 \\ 0 & -.022375 & -.016781 \\ -.022375 & 0 & +.010627 \\ -.016781 & +.010627 & 0 \end{bmatrix}' \tag{2.28}$$

This tells us that there are 5 difference dimensions in the space spanned by the differences in mean, variance, and covariance in our population. Since we only have three predictors, some of the dimensions are redundant and linear combinations of the others. Examining the eigenvalues of the kernel matrix $\mathbf{SpSp}'$ yields the following.

$$\begin{aligned} \lambda_1 &= 1.0017336139 \\ \lambda_2 &= 0.0014197888 \\ \lambda_3 &= 0.0001094204 \end{aligned} \tag{2.29}$$

The third eigenvalue is very small, but it is not zero. In practice, with the tests that we will develop later in this document and sufficiently large sample sizes, we would reject the hypothesis that $\lambda_3 = 0$ and that there are at most 2 SMVCIR dimensions.

We would conclude that there is covariance difference, in addition to the variance and mean differences. This is not true. Using the square-root of the inverse of the marginal variance leads to an extra SMVCIR difference dimension and misleads our analysis of the population.

Earlier we ignored the group weights since they were equal. To formally get the SMVCIR spanning set, we scale the spanning vectors corresponding with each group by the square root of that group's population proportion. Each spanning set vector represents the difference of a particular group from the overall group average. This weighting gives a greater magnitude to populous group spanning vectors than non-populous group spanning vectors. So the effect of a difference with a large group is

shown more strongly than that of a small group in the SMVCIR space.

If we examine this scaling from the population perspective, treating the scaled vectors as again-transformed means and variances, it seems that the Strong Conversion Principle is violated for this new post-standardization transformation.

But the SMVCIR spanning set already contains the differences in means, variances, and covariances of the groups. And these differences, not the original population parameters are being scaled. A standardization transformation may only be applied to data from the population. When we perform this rescaling of the centered parameters, there is no population standardization transformation analog to the rescaling. The variance matrices have already been centered, resulting in potentially negative variance components, so no population can correspond to the centered parameters that are being rescaled.

So application of the Strong Conversion Principle to the weighting transformation is erroneous. Additionally, the weighting transformation maintains a form of invariance for the SMVCIR spanning set. The SMVCIR space of differences in mean, variance, and covariances is formed from linear combinations of the vectors in the SMVCIR spanning set. So altering the SMVCIR spanning set by multiplying its component vectors by certain scalars will result in the exact same SMVCIR space.

Now we will re-perform the SMVCIR analysis using the correct standardization in equations (2.15) - (2.18).

The pooled variance from (2.17) and (2.18) is given by the following.

$$\begin{array}{l}\sum_{i=1}^{g} p_i \mathbf{\Sigma_i} = \\ .5 \begin{bmatrix} 10001.1^{-1} & 0 & 0 \\ 0 & 10001^{-1} & 0 \\ 0 & 0 & 5626^{-1} \end{bmatrix} + \\ .5 \begin{bmatrix} 1.2/10001.1 & 0 & 0 \\ 0 & 10001^{-1} & 0 \\ 0 & 0 & 5626^{-1} \end{bmatrix} = \\ \begin{bmatrix} 2.2/20002.2 = 1.1/10001.1 & 0 & 0 \\ 0 & 10001^{-1} & 0 \\ 0 & 0 & 5626^{-1} \end{bmatrix}\end{array} \tag{2.30}$$

This leads to the following centered variances.

$$\begin{array}{l}\mathbf{\Sigma_1^{zc}} = \\ \begin{bmatrix} 10001.1^{-1} & 0 & 0 \\ 0 & 10001^{-1} & 0 \\ 0 & 0 & 5626^{-1} \end{bmatrix} - \sum_{i=1}^{g} p_i \mathbf{\Sigma_i} = \\ \begin{bmatrix} -.1/10001.1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\end{array} \tag{2.31}$$

$$\begin{array}{l}\mathbf{\Sigma_2^{zc}} = \\ \begin{bmatrix} 1.2/10001.1 & 0 & 0 \\ 0 & 10001^{-1} & 0 \\ 0 & 0 & 5626^{-1} \end{bmatrix} - \sum_{i=1}^{g} p_i \mathbf{\Sigma_i} = \\ \begin{bmatrix} .1/10001.1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\end{array} \tag{2.32}$$

This leads to the following SMVCIR spanning set. Again, for clarity, we ignore

the weights since both groups have the same prevalence.

$$\mathbf{Sp}' \approx \begin{bmatrix} -.99995 & -.99995 & -.99991 \\ +.99995 & +.99995 & +.99991 \\ -.1/10001.1 & 0 & 0 \\ +.1/10001.1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}' \tag{2.33}$$

This matrix clearly has rank 2. As we hoped, one difference dimension is found for the means and another for the variances. There is no covariance difference.

D. Second Example, Means and Variances Similar Magnitudes

Let us do an example where the means and variance differences are of more similar magnitudes. We have three groups with the following original scale parameters.

$$\begin{aligned} &p_1 = p_2 = .25 \\ &p_3 = .5 \\ &\boldsymbol{\mu}_1' = \begin{bmatrix} 10 & 1 & -10 & 16 & 3 \end{bmatrix}' \\ &\boldsymbol{\mu}_2' = \boldsymbol{\mu}_3' = \begin{bmatrix} 1 & 20 & 3 & 4 & 50 \end{bmatrix}' \end{aligned} \tag{2.34}$$

$$\begin{aligned} &\boldsymbol{\Sigma}_1 = \begin{bmatrix} 10 & 6 & 0 & 0 & 0 \\ 6 & 10 & 0 & 0 & 0 \\ 0 & 0 & 10 & 0 & 0 \\ 0 & 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 0 & 10 \end{bmatrix} \\ &\boldsymbol{\Sigma}_2 = 10 * \mathbf{I}_5 \\ &\boldsymbol{\Sigma}_3 = \mathbf{I}_5 \end{aligned} \tag{2.35}$$

The SMVCIR algorithm should form a mean difference dimension, a single variance difference dimension, and two covariance difference dimensions corresponding to

the first and second columns of $\mathbf{\Sigma_1}$. These differences are unique in scale, so the rank of the SMVCIR spanning set should be 4.

This is the marginal mean.

$$\begin{aligned} \boldsymbol{\mu}' &= \left(\textstyle\sum_{i=1}^{g} p_i \boldsymbol{\mu}_\mathbf{i}\right)' \\ &= \begin{bmatrix} 3.25 & 15.25 & -.25 & 7 & 38.25 \end{bmatrix}' \end{aligned} \tag{2.36}$$

This is the marginal variance.

$$\mathbf{\Sigma} \approx \begin{bmatrix} +20.6875 & -30.5625 & -21.9375 & +20.2500 & -79.3125 \\ -30.5625 & +73.1875 & +46.3125 & -42.7500 & +167.4380 \\ -21.9375 & +46.3125 & +37.1875 & -29.2500 & +114.5625 \\ +20.2500 & -42.7500 & -29.2500 & +32.5000 & -105.7500 \\ -79.3125 & +167.4375 & +114.5625 & -105.7500 & +419.6875 \end{bmatrix} \tag{2.37}$$

The inverse square root of this matrix is the following.

$$\mathbf{\Sigma}^{-\frac{1}{2}} \approx \begin{bmatrix} +.4194565 & -.0280865 & +.0201484 & -.0185985 & +.0728443 \\ -.0280865 & +.3828199 & -.0346112 & +.0319488 & -.1251326 \\ +.0201484 & -.0346112 & +.4049720 & +.0197811 & -.0774758 \\ -.0185985 & +.0319488 & +.0197811 & +.4081420 & +.0715161 \\ +.0728443 & -.1251326 & -.0774758 & +.0715161 & +.1462967 \end{bmatrix} \tag{2.38}$$

This is the inverse square root of the marginal variance diagonal.

$$\operatorname{diag}\left(\mathbf{\Sigma}\right)^{-\frac{1}{2}} \approx \begin{bmatrix} 0.2198599 & 0 & 0 & 0 & 0 \\ 0 & 0.1168911 & 0 & 0 & 0 \\ 0 & 0 & 0.1639840 & 0 & 0 \\ 0 & 0 & 0 & 0.1754116 & 0 \\ 0 & 0 & 0 & 0 & 0.0488132 \end{bmatrix} \tag{2.39}$$

These are the centered means.

$$\begin{aligned} \boldsymbol{\mu}_\mathbf{1}^{\mathbf{c}\prime} &= \begin{bmatrix} 6.75 & -14.25 & -9.75 & 9 & -35.25 \end{bmatrix}' \\ \boldsymbol{\mu}_\mathbf{2}^{\mathbf{c}\prime} = \boldsymbol{\mu}_\mathbf{3}^{\mathbf{c}\prime} &= \begin{bmatrix} -2.25 & 4.75 & 3.25 & -3 & 11.75 \end{bmatrix}' \end{aligned} \tag{2.40}$$

First we will standardize with the marginal variance in (2.38) and show how this

is problematic. This leads to the following standardized centered means.

$$\begin{aligned}
&\boldsymbol{\mu}_{\mathbf{1}}^{\mathbf{cz'}} \approx \\
&\begin{bmatrix} +.2999697 & -.6088451 & -.4102155 & +.3786605 & -1.4830869 \end{bmatrix}' \\
&\boldsymbol{\mu}_{\mathbf{2}}^{\mathbf{cz'}} = \boldsymbol{\mu}_{\mathbf{3}}^{\mathbf{cz'}} \approx \\
&\begin{bmatrix} -.0999900 & +.2029484 & +.1367385 & -.1262202 & +.4943623 \end{bmatrix}'
\end{aligned} \tag{2.41}$$

These are the standardized variances.

$$\begin{aligned}
&\boldsymbol{\Sigma}_{\mathbf{1}}^{\mathbf{z}} \approx \\
&\begin{bmatrix} +1.686535 & +0.638792 & +0.025212 & -0.023273 & +0.091152 \\ +0.638792 & +1.523143 & -0.122944 & +0.113486 & -0.444489 \\ +0.025212 & -0.122944 & +1.711631 & +0.098354 & -0.385221 \\ -0.023273 & +0.113486 & +0.098354 & +1.727393 & 0.355589 \\ +0.091151 & -0.444489 & -0.385221 & +0.355589 & +0.425460 \end{bmatrix}
\end{aligned} \tag{2.42}$$

$$\begin{aligned}
&\boldsymbol{\Sigma}_{\mathbf{2}}^{\mathbf{z}} \approx \\
&\begin{bmatrix} +1.827908 & -0.329399 & +0.115715 & -0.106814 & +0.418353 \\ -0.329399 & +1.652168 & -0.175056 & +0.161590 & -0.632893 \\ +0.115715 & -0.175056 & +1.720000 & +0.090630 & -0.354966 \\ -0.106814 & +0.161590 & +0.090630 & +1.734524 & 0.327661 \\ +0.418353 & -0.632893 & -0.354966 & +0.327661 & +0.534842 \end{bmatrix}
\end{aligned} \tag{2.43}$$

$$\begin{aligned}
&\boldsymbol{\Sigma}_{\mathbf{3}}^{\mathbf{z}} \approx \\
&\begin{bmatrix} +.182791 & -.032940 & +.011571 & -.010681 & +.041835 \\ -.032940 & +.165217 & -.017506 & +.016159 & -.063289 \\ +.011572 & -.017506 & +.172000 & +.009063 & -.035497 \\ -.010681 & +.016159 & +.009063 & +.173452 & +.032766 \\ +.041835 & -.063289 & -.035497 & +.032766 & +.053484 \end{bmatrix}
\end{aligned} \tag{2.44}$$

The pooled standardized variance is given by the following.

$$\begin{aligned}
&\textstyle\sum_{i=1}^{g} p_i \boldsymbol{\Sigma}_{\mathbf{i}} \approx \\
&\begin{bmatrix} +.970006 & +.060878 & +.041017 & -.037862 & +.148294 \\ +.060878 & +.876436 & -.083253 & +.076849 & -.300990 \\ +.041017 & -.083253 & +.943908 & +.051778 & -.202795 \\ -.037862 & +.076849 & +.051778 & +.952205 & +.187196 \\ +.148294 & -.300990 & -.202795 & +.187196 & +.266818 \end{bmatrix}
\end{aligned} \tag{2.45}$$

This leads to the following centered standardized variances.

$$\boldsymbol{\Sigma}_{\mathbf{1}}^{\mathbf{zc}} \approx \begin{bmatrix} +.716529 & +.577914 & -.015806 & +.014590 & -.057143 \\ +.577914 & +.646707 & -.039691 & +.036638 & -.143498 \\ -.015805 & -.039691 & +.767724 & +.046577 & -.182426 \\ +.014590 & +.036638 & +.046577 & +.775188 & +.168393 \\ -.057143 & -.143498 & -.182426 & +.168393 & +.158642 \end{bmatrix} \tag{2.46}$$

$$\boldsymbol{\Sigma}_{\mathbf{2}}^{\mathbf{zc}} \approx \begin{bmatrix} +.857902 & -.390277 & +.074697 & -.068951 & +.270059 \\ -.390277 & +.775732 & -.091803 & +.084741 & -.331903 \\ +.074697 & -.091803 & +.776092 & +.038852 & -.152171 \\ -.068951 & +.084741 & +.038852 & +.782318 & +.140466 \\ +.270059 & -.331903 & -.152171 & +.140466 & +.268025 \end{bmatrix} \tag{2.47}$$

$$\boldsymbol{\Sigma}_{\mathbf{3}}^{\mathbf{zc}} \approx \begin{bmatrix} -.787215 & -.093818 & -.029446 & +.027181 & -.1064584 \\ -.093818 & -.711219 & +.065747 & -.060690 & +.2377007 \\ -.029446 & +.065747 & -.771908 & -.042715 & +.1672985 \\ +.027181 & -.060690 & -.042715 & -.778753 & -.1544294 \\ -.106458 & +.237701 & +.167299 & -.154429 & -.2133335 \end{bmatrix} \tag{2.48}$$

Now we can give the SMVCIR spanning set matrix. Here we use the weights. The group proportions are different, so we will include them in the spanning set to be formal and clear.

$$\mathbf{Sp}' \approx$$

$$\left[\begin{array}{ll}
4^{-.5} & \left[\begin{array}{ccccc} +0.299970 & -0.608845 & -0.410216 & +0.378661 & -1.483087 \end{array}\right] \\
4^{-.5} & \left[\begin{array}{ccccc} -0.099990 & +0.202948 & +0.136739 & -0.126220 & +0.494362 \end{array}\right] \\
2^{-.5} & \left[\begin{array}{ccccc} -0.099990 & +0.202948 & +0.136739 & -0.126220 & +0.494362 \end{array}\right] \\
4^{-.5} & \left[\begin{array}{ccccc} +0.716529 & +0.646707 & +0.767724 & +0.775178 & +0.158642 \end{array}\right] \\
4^{-.5} & \left[\begin{array}{ccccc} +0.857902 & +0.775732 & +0.776092 & +0.782318 & +0.268025 \end{array}\right] \\
2^{-.5} & \left[\begin{array}{ccccc} -0.787215 & -0.711219 & -0.771908 & -0.778753 & -0.213334 \end{array}\right] \\
 & \left[\begin{array}{ccccc} 0 & -0.577914 & -0.015805 & -0.014590 & -0.057143 \\ -0.577914 & 0 & -0.039691 & +0.036638 & -0.143498 \\ -0.015805 & -0.039691 & 0 & +0.046577 & -0.182426 \\ +0.014590 & +0.036638 & +0.046577 & 0 & +0.168393 \\ -0.057143 & -0.143498 & -0.182426 & +0.168393 & 0 \end{array}\right] \\
4^{-.5} & \left[\begin{array}{ccccc} 0 & -0.390277 & +0.074697 & -0.068951 & +0.270059 \\ -0.390277 & 0 & -0.091803 & +0.084741 & -0.331903 \\ +0.074697 & -0.091803 & 0 & +0.038852 & -0.152171 \\ -0.068951 & +0.084741 & +0.038852 & 0 & +0.140466 \\ +0.270059 & -0.331903 & -0.152171 & +0.140466 & 0 \end{array}\right] \\
 & \left[\begin{array}{ccccc} 0 & -0.093818 & -0.029446 & +0.027181 & -0.106458 \\ -0.093818 & 0 & +0.065747 & -0.060690 & +0.237701 \\ -0.029446 & +0.065747 & 0 & -0.042715 & +0.167299 \\ +0.027181 & -0.060690 & -0.042715 & 0 & -0.154429 \\ -0.106458 & +0.237701 & +0.167299 & -0.154429 & 0 \end{array}\right] \\
2^{-.5} &
\end{array}\right]' \tag{2.49}$$

Performing the scalar multiplication we obtain

$$\mathbf{Sp}' \approx \begin{bmatrix} +.149985 & -.304423 & -.205108 & +.189330 & -.741544 \\ -.049995 & +.101474 & +.068369 & -.063110 & +.247181 \\ -.070704 & +.143506 & +.096689 & -.089251 & +.349567 \\ +.358265 & +.323353 & +.383862 & +.387594 & +.079321 \\ +.428951 & +.387866 & +.388046 & +.391159 & +.134012 \\ -.556645 & -.502908 & -.545821 & -.550662 & -.150850 \\ 0 & +.288957 & -.007903 & +.007295 & -.028571 \\ +.288957 & 0 & -.019846 & +.018319 & -.071749 \\ -.007903 & -.019846 & 0 & +.023288 & -.091213 \\ +.007295 & +.018319 & +.023288 & 0 & +.084197 \\ -.028571 & -.071749 & -.091213 & +.084197 & 0 \\ 0 & -.195139 & +.037349 & -.034476 & +.135030 \\ -.195139 & 0 & -.045902 & +.042371 & -.165952 \\ +.037349 & -.045902 & 0 & +.019426 & -.076086 \\ -.034476 & +.042371 & +.019426 & 0 & +.070233 \\ +.135030 & -.165952 & -.076086 & +.070233 & 0 \\ 0 & -.066340 & -.020821 & +.019220 & -.075278 \\ -.066340 & 0 & +.046490 & -.042914 & +.168080 \\ -.020821 & +.046490 & 0 & -.030204 & +.118298 \\ +.019220 & -.042914 & -.030204 & 0 & -.109198 \\ -.075278 & +.168080 & +.118298 & -.109198 & 0 \end{bmatrix}' \tag{2.50}$$

For the first 3 vectors, the centered means, we see that the second and third vectors are scalar multiples of each other. We can actually show that each of the centered means is a scalar multiple of the same vector.

The two unique differences are linearly independent. So the mean differences add one unique difference dimension to the SMVCIR space. In practice we don't have to do this specific calculation ourselves, the eigen decomposition will do it automatically. We do it here to be instructive.

We start with the transposed mean differences, the first row corresponds to the first mean, etc.

$$\begin{matrix} 4^{-.5} \\ 4^{-.5} \\ 2^{-.5} \end{matrix} \begin{bmatrix} +0.299970 & -0.608845 & -0.410216 & +0.378661 & -1.483087 \\ -0.099990 & +0.202948 & +0.136739 & -0.126220 & +0.494362 \\ -0.099990 & +0.202948 & +0.136739 & -0.126220 & +0.494362 \end{bmatrix} \tag{2.51}$$

Ignoring the scaling factors, we divide the components of the first vector by the second. Each component has a scaling factor of -3 . So we can rewrite the means as the following.

$$\begin{array}{l} 4^{-.5}\left[\begin{array}{ccccc} +0.299970 & -0.608845 & -0.410216 & +0.378661 & -1.483087 \end{array}\right] \\ 4^{-.5}\left[\begin{array}{ccccc} -0.099990 & +0.202948 & +0.136739 & -0.126220 & +0.494362 \end{array}\right] \\ 2^{-.5}\left[\begin{array}{ccccc} -0.099990 & +0.202948 & +0.136739 & -0.126220 & +0.494362 \end{array}\right] \\ \approx \\ 4^{-.5} * -3\left[\begin{array}{ccccc} -0.099990 & +0.202948 & +0.136739 & -0.126220 & +0.494362 \end{array}\right] \\ 4^{-.5}\left[\begin{array}{ccccc} -0.099990 & +0.202948 & +0.136739 & -0.126220 & +0.494362 \end{array}\right] \\ 2^{-.5}\left[\begin{array}{ccccc} -0.099990 & +0.202948 & +0.136739 & -0.126220 & +0.494362 \end{array}\right] \\ \approx \\ \left[\begin{array}{ccccc} +0.149985 & -0.304423 & -0.205108 & +0.189330 & -0.741544 \\ -0.049995 & +0.101474 & +0.068370 & -0.063110 & +0.247181 \\ -0.070704 & +0.143506 & +0.096689 & -0.089251 & +0.349567 \end{array}\right] \end{array} \tag{2.52}$$

Now we look at the variances. We do component division on group 1 versus group 2, this involves the group 1 variance for predictor 1 divided by the group 2 variance for predictor 1 and so on.

$$\begin{array}{l} 4^{-.5}\left[\begin{array}{ccccc} .716529 & .646707 & .767724 & .775188 & .158642 \end{array}\right] \\ / \\ 4^{-.5}\left[\begin{array}{ccccc} .857902 & .775732 & .776092 & .782318 & .268025 \end{array}\right] \\ \approx \\ \left[\begin{array}{ccccc} .835211 & .833673 & .989217 & .990886 & .591895 \end{array}\right] \end{array} \tag{2.53}$$

The component wise division does not lead to single factor, like the -3 factor that we saw for the means. This confirms that there are multiple linearly independent vectors among the centered variances. This finding contradicts our conclusions after (2.35). The first two groups have the same variances, differing from those of the third by a factor of 10. The strong conversion of the variance parameters is violated by this standardization method, and as a result we have extra dimensions for the SMVCIR space. To conclude the example using the wrong standardization, we compute the

eigenvalues of the kernel, $\mathbf{SpSp'}$. The number of non-zero eigenvalues is equal to number of dimensions in the SMVCIR, or the unique difference between the groups with respect to means, variances, and covariances. None of the five are zero. We expected 4 dimensions for the SMVCIR space, and so only four of the eigenvalues should be non-zero.

$$\begin{aligned}\lambda_1 &= 2.51707729\\ \lambda_2 &= 1.09418160\\ \lambda_3 &= 0.20701958\\ \lambda_4 &= 0.07181876\\ \lambda_5 &= 0.03526660\end{aligned} \tag{2.54}$$

Now we will see what happens when we use the square-root inverse of the diagonal of the marginal variance to standardize. First we re-calculate the standardized centered means.

$$\begin{aligned}&\boldsymbol{\mu}_{\mathbf{1}}^{\mathbf{cz'}} \approx\\ &\begin{bmatrix} +1.484054 & -1.665699 & -1.598844 & +1.578704 & -1.720664 \end{bmatrix}'\\ &\boldsymbol{\mu}_{\mathbf{2}}^{\mathbf{cz'}} = \boldsymbol{\mu}_{\mathbf{3}}^{\mathbf{cz'}} \approx\\ &\begin{bmatrix} -0.494685 & +0.555233 & +0.532948 & -0.526235 & +0.573555 \end{bmatrix}'\end{aligned} \tag{2.55}$$

Now we recalculate the standardized variances.

$$\boldsymbol{\Sigma}_{\mathbf{1}}^{\mathbf{z}} \approx \begin{bmatrix} 0.483384 & 0.154198 & 0 & 0 & 0\\ 0.154198 & 0.136635 & 0 & 0 & 0\\ 0 & 0 & 0.268908 & 0 & 0\\ 0 & 0 & 0 & 0.307692 & 0\\ 0 & 0 & 0 & 0 & 0.023827 \end{bmatrix} \tag{2.56}$$

$$\boldsymbol{\Sigma}_{\mathbf{2}}^{\mathbf{z}} \approx \begin{bmatrix} 0.483384 & 0 & 0 & 0 & 0\\ 0 & 0.136635 & 0 & 0 & 0\\ 0 & 0 & 0.268908 & 0 & 0\\ 0 & 0 & 0 & 0.307692 & 0\\ 0 & 0 & 0 & 0 & 0.023827 \end{bmatrix} \tag{2.57}$$

$$\mathbf{\Sigma_3^z} \approx \begin{bmatrix} 0.048338 & 0 & 0 & 0 & 0 \\ 0 & 0.013664 & 0 & 0 & 0 \\ 0 & 0 & 0.026891 & 0 & 0 \\ 0 & 0 & 0 & 0.030769 & 0 \\ 0 & 0 & 0 & 0 & 0.002383 \end{bmatrix} \tag{2.58}$$

And then we center them.

$$\mathbf{\Sigma_1^{zc}} \approx \begin{bmatrix} +0.217523 & +0.115649 & 0 & 0 & 0 \\ +0.115649 & +0.061486 & 0 & 0 & 0 \\ 0 & 0 & +0.121008 & 0 & 0 \\ 0 & 0 & 0 & +0.138462 & 0 \\ 0 & 0 & 0 & 0 & +0.010722 \end{bmatrix} \tag{2.59}$$

$$\mathbf{\Sigma_2^{zc}} \approx \begin{bmatrix} +0.217523 & -0.038550 & 0 & 0 & 0 \\ -0.038550 & +0.061486 & 0 & 0 & 0 \\ 0 & 0 & +0.121008 & 0 & 0 \\ 0 & 0 & 0 & +0.138462 & 0 \\ 0 & 0 & 0 & 0 & +0.010722 \end{bmatrix} \tag{2.60}$$

$$\mathbf{\Sigma_3^{zc}} \approx \begin{bmatrix} -0.217523 & -0.038550 & 0 & 0 & 0 \\ -0.038550 & -0.061486 & 0 & 0 & 0 \\ 0 & 0 & -0.121008 & 0 & 0 \\ 0 & 0 & 0 & -0.138462 & 0 \\ 0 & 0 & 0 & 0 & -0.010722 \end{bmatrix} \tag{2.61}$$

This leads to the following SMVCIR spanning set.

$$
\mathbf{Sp}' \approx
\begin{bmatrix}
4^{-.5}\begin{bmatrix} +1.484054 & -1.665699 & -1.598844 & +1.578704 & -1.720664 \end{bmatrix} \\
4^{-.5}\begin{bmatrix} -0.494685 & +0.555233 & +0.532948 & -0.526235 & +0.573555 \end{bmatrix} \\
2^{-.5}\begin{bmatrix} -0.494685 & +0.555233 & +0.532948 & -0.526235 & +0.573555 \end{bmatrix} \\
4^{-.5}\begin{bmatrix} +0.217523 & +0.061486 & +0.121008 & +0.138462 & +0.010722 \end{bmatrix} \\
4^{-.5}\begin{bmatrix} +0.217523 & +0.061486 & +0.121008 & +0.138462 & +0.010722 \end{bmatrix} \\
2^{-.5}\begin{bmatrix} -0.217523 & -0.061486 & -0.121008 & -0.138462 & -0.010722 \end{bmatrix} \\
4^{-.5}\begin{bmatrix} 0 & +0.115649 & 0 & 0 & 0 \\ +0.115649 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \\
4^{-.5}\begin{bmatrix} 0 & -0.038550 & 0 & 0 & 0 \\ -0.038550 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \\
2^{-.5}\begin{bmatrix} 0 & -0.038550 & 0 & 0 & 0 \\ -0.038550 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}
\end{bmatrix}'
\tag{2.62}
$$

A component-wise division of the first means difference vector by the second shows the two linked again linked by a -3 factor. So we get one unique mean difference. The linear relation of all three variance vectors is very obvious, and we obtain one variance difference.

This difference is linearly independent from our mean difference. We can see this by looking at the signs of each element.

We note that $0.115649 = -3 * -0.038550$, so we obtain two covariance differences, one for each column of the predictor covariance matrix that has a $(1, 2)$ entry. These differences are linearly independent of each other, and of the mean and variance difference.

So we should have 4 unique differences, and 4 dimensions in the SMVCIR space.

We perform an eigen decomposition of $\mathbf{SpSp'}$ and expect to see 4 non-zero eigenvalues.

$$\begin{aligned} \lambda_1 &= 4.334515000 \\ \lambda_2 &= 0.083501560 \\ \lambda_3 &= 0.003049704 \\ \lambda_4 &= 0.001630818 \\ \lambda_5 &= 0 \end{aligned} \tag{2.63}$$

Our calculation engine for this report, the statistical programming language R (created and maintained by R Development Core Team (2009)), actually reported $5.717472 * 10^{-18}$ for the last eigenvalue. This non-zero value is very close to zero, and believed to be a result of the finite precision involved in computing the eigen decomposition numerically.

E. Final Cautions

We have concluded our examples. We have thoroughly demonstrated how using the Mahalanobis transformation may violate the Strong Conversion Principle and introduce additional difference dimensions to the SMVCIR space. It will correct marginal linear dependence problems however. We will now see how this happens and how the scaling transformation does not correct marginal linear dependence.

At the final stage of the SMVCIR algorithm, the standardized predictor coordinates are translated to the SMVCIR space.

Standardized predictor observation $\mathbf{z_i}$, has these SMVCIR coordinates.

$$\mathbf{s_i} = \mathbf{E'z_i} = \begin{bmatrix} \mathbf{e_1} & \cdots & \mathbf{e_d} & \cdots & \mathbf{e_k} \end{bmatrix}' \mathbf{z_i} \tag{2.64}$$

The $\mathbf{e_i}$ vectors are the eigenvectors of the SMVCIR kernel calculated using the standardized predictors. The number d is the dimension of the SMVCIR space. In the transformed $\mathbf{s_i}$ observation we ignore the last $k - d$ components. We only write the transformation in this manner, including all k elements, so that we may more easily explain the resulting variance matrix of the transformed variables. It is not necessarily diagonal.

$$\begin{aligned} &\mathrm{Var}\left[\mathbf{s_i}\right] = \\ &\mathbf{E}'\mathrm{Var}\left[\mathbf{z_i}\right]\mathbf{E} \end{aligned} \tag{2.65}$$

The $\mathbf{E}$ matrix is orthogonal (actually orthonormal), so when $\mathrm{Var}\left[\mathbf{z_i}\right]$ is the identity, $\mathrm{Var}\left[\mathbf{s_i}\right]$ is the identity. Outside of this case, the orthogonality of $\mathrm{Var}\left[\mathbf{s_i}\right]$ is not assured. By standardizing using the Mahalanobis transformation, where we multiple the original predictors by the square-root inverse of the marginal variance, we ensure that the standardized predictors have an identity variance/covariance matrix. So under the Mahalanobis standardization, $\mathrm{Var}\left[\mathbf{s_i}\right]$ is the identity. So $\mathrm{Var}\left[\mathbf{s_i}\right]$ is diagonal under Mahalanobis standardization.

When $\mathrm{Var}\left[\mathbf{s_i}\right]$ is not diagonal, it means that the SMVCIR coordinates are correlated. So visualization and interpretation of the SMVCIR dimensions may be difficult. Patterns between the dimensions will be distorted in the actual data displayed because the coordinates involved in the patterns are correlated.

On page 102 of Cook (1998), it was suggested that orthogonal coordinates be used to assess the features of the central subspace of the predictors on the response. So linear combinations of the predictors that are linearly independent of each other are advocated. But Dr. Cook went on to say in Cook (1998), "Interpretation may be facilitated at other times by letting the columns of A correspond to linear combinations of particular interest, provided the latter are not too strongly correlated."

He was discussing the choice of vectors to multiply the data by to obtain the final transformed coordinates. So the final translated coordinates (linear combinations of the predictors) may be formulated to correspond to difference relationships that are of interest, so long as they are not too strongly correlated.

We take a similar approach here. Examination of how groups differ in mean, variance, and covariance is easier when we are not worried about the plotting dimensions being very dependent and distorting our perception. But it is not necessarily worth standardizing with $\mathbf{\Sigma}^{-.5}$ in the Mahalanobis transformation (potentially violating the Strong Conversion Principle and causing problems like those we documented in the previous examples) to safeguard this perception.

In practice, we suggest checking the marginal correlations of the final SMVCIR coordinates prior to and in conjunction with examining the plots and/or regression of the coordinates on the standardized predictors. If they are within acceptable tolerances, then analysis of plots and the regressions may be carried out without extra caution. If some are large, then those dimensions should be handled carefully. Differences and patterns may still be detectable, but their appearance is distorted by the correlations.

To be conservative, it may be best to return to the original standardization $\mathbf{\Sigma}^{-.5}$ instead of trying to interpret the SMVCIR results when the final SMVCIR coordinates are highly correlated. One should be careful and alert to noise dimensions in this case.

We should mention that the other predictor problem, marginal scaling differences, is solved by both the scaling and Mahalanobis transformation. Moreover under the scaling standardization, the small scale of $\text{Var}\left[\mathbf{s_i}\right]$ is ensured by the orthonormality of the eigenvectors and the correspondence of $\text{Var}\left[\mathbf{z_i}\right] = \text{Corr}\left[\mathbf{z_i}\right] = \text{Corr}\left[\mathbf{x_i}\right]$. Since the covariance matrix of the standardized predictors is equivalent to its correlation matrix, all the predictors have the same scale.

In practice, one should use the sample mean for estimating $\boldsymbol{\mu}$ and the sample variances for estimating the diagonal elements of $\boldsymbol{\Sigma}$. Our dimensionality test is developed under these assumptions. Using these statistics, we have strong consistent estimators of the mean and variance parameters. Taking the square-root inverse of the $\boldsymbol{\Sigma}$ diagonal is a continuous operation, so applying the operation to the sample estimates of the variances will provide a strong consistent estimate of diag $\left(\mathbf{S}^{-\frac{1}{2}}\right)$. So as before, we suggest standardizing the population with $\mathbf{z_i} = \text{diag}\left(\mathbf{S}^{-\frac{1}{2}}\right)(\mathbf{x_i} - \bar{\mathbf{x}})$. If the resulting SMVCIR coordinates are too highly correlated you may use $\mathbf{z_i} = \mathbf{S}^{-\frac{1}{2}}(\mathbf{x_i} - \bar{\mathbf{x}})$. Watch out for noise dimensions and be wary of switched relationships among the differences if you perform the Mahalanobis standardization. Particularly strong differences should show through regardless of which standardization you use. Note that our dimensionality test is not developed for use under this alternative standardization method.

Now we will develop the test of the dimension of the SMVCIR space.

CHAPTER III

DIMENSIONALITY TEST: SPANNING SET

By determining the dimension of the SMVCIR space, d, we decide how many of the transformed variables need to be examined to see the mean, variance, and covariance differences between groups.

Our dimensionality test will allow us to infer on the number of dimensions. The approach we take was inspired by that of the technical report Yin (2005). Yin's test hinges on the data coming from a standardized distribution. Our data will come from a non-standardized distribution and we must make it standardized, adding an extra level of complexity.

The dimension of the SMVCIR space is the same as the rank of the spanning set matrix. This is the same as the number of non-zero singular values of the spanning set $\mathbf{K}$ and non-zero eigenvalues of the kernel $\mathbf{KK}'$. Using the eigen and singular value decompositions described in Chapter I, we can diagonalize these matrices and obtain the eigen and singular values.

We suggest a one-sided alternative test. The eigen and singular values are always non-negative, so there is no obvious point of symmetry to base a two sided hypothesis test around.

To test the null hypothesis $d = i$ versus the alternative $d > i$, we could sum the $d+1, \ldots, k$ diagonal elements of one of the diagonalized matrices just mentioned, and reject the null if the resultant value is too large.

To actually estimate a value of d, we use an iterative series of tests.

$$\begin{aligned}
&\text{Test } 0 : \mathrm{H}_0 : d = 0 \text{ vs } \mathrm{H}_1 : d > 0 \\
&\text{Test } 1 : \mathrm{H}_0 : d = 1 \text{ vs } \mathrm{H}_1 : d > 1 \\
&\vdots \\
&\text{Test i} : \mathrm{H}_0 : d = i \text{ vs } \mathrm{H}_1 : d > i \\
&\vdots \\
&\text{Test k-1} : \mathrm{H}_0 : d = k - 1 \text{ vs } \mathrm{H}_1 : d = k
\end{aligned} \tag{3.1}$$

Note that there are k tests. We have $\text{rank}\,(\mathbf{K}) \leq k$ because of the inclusion of the covariance matrix columns.

We start with test 0. If we reject the null, then we perform test 1. If we reject the null again, we move to the third test, etc. We move through the tests until we finally accept a null hypothesis or reach the last test. Our accepted value of d is the index of the last test we performed, or k if we reject the null hypothesis of the last test.

For test i, we calculate the statistic

$$\hat{\Lambda}_i = n \sum_{j=i+1}^{k} \hat{\lambda}_j \tag{3.2}$$

The scalar values $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_k$ are the eigenvalues of the estimated kernel matrix $\mathbf{K_n K_n'}$. To perform test i, we have to determine the distribution of $\hat{\Lambda}_i$ under $\mathrm{H}_0 : d = i$. The same statistic and a similar testing scheme is used in SAVE (Cook and Yin (2001)) and SIR (Li (1991)).

In this chapter, we develop asymptotic distributional results for our estimate of $\mathbf{K}$ (1.15). In the next chapter we use these to develop asymptotic distributional results for the final test statistic $\hat{\Lambda}_i$. This approach has the disadvantage of yielding only approximate results, but it has the advantage of complete generality. We will

use only central limit theorems and basic convergence theorems, so our results will be approximately valid regardless of the underlying distributions of the predictors and response.

The standardization that we use will affect the derivation of our test of dimensionality. As we advocated the scaling transformation previously, we will develop our dimensionality test based on this standardization.

The structure of this chapter can be more easily explained using a schematic, which we draw in Figure 5. We will devote a section to each "node" in the schematic. These represent stages in the calculation of the spanning set estimate $\mathbf{K_n}$. We begin with a central limit theorem invoked on group pseudo-first and pseudo-second moments, group proportions, and marginal first and second moments. Then we invoke a delta method with 5 separate stages. In the first stage group first and second moments are formed using the group proportions and pseudo moments from the central limit theorem stage. In the next stage, we calculate the group/marginal variances/covariances using the second and first moments. Then we finally standardize the group means and variances using the marginal means and variances. This forms mean differences as well, as we subtract out the marginal mean. In the next stage, we create variance and covariance differences by subtracting out the pooled variance/covariance matrix. Following this, in the final delta method stage, we weight the calculated group mean, variance, and covariance differences using the group proportions. This ends the delta method. To obtain the final spanning set distribution we then use a permutation matrix to stack the variances and place zeroes across the diagonal of the covariance matrices.

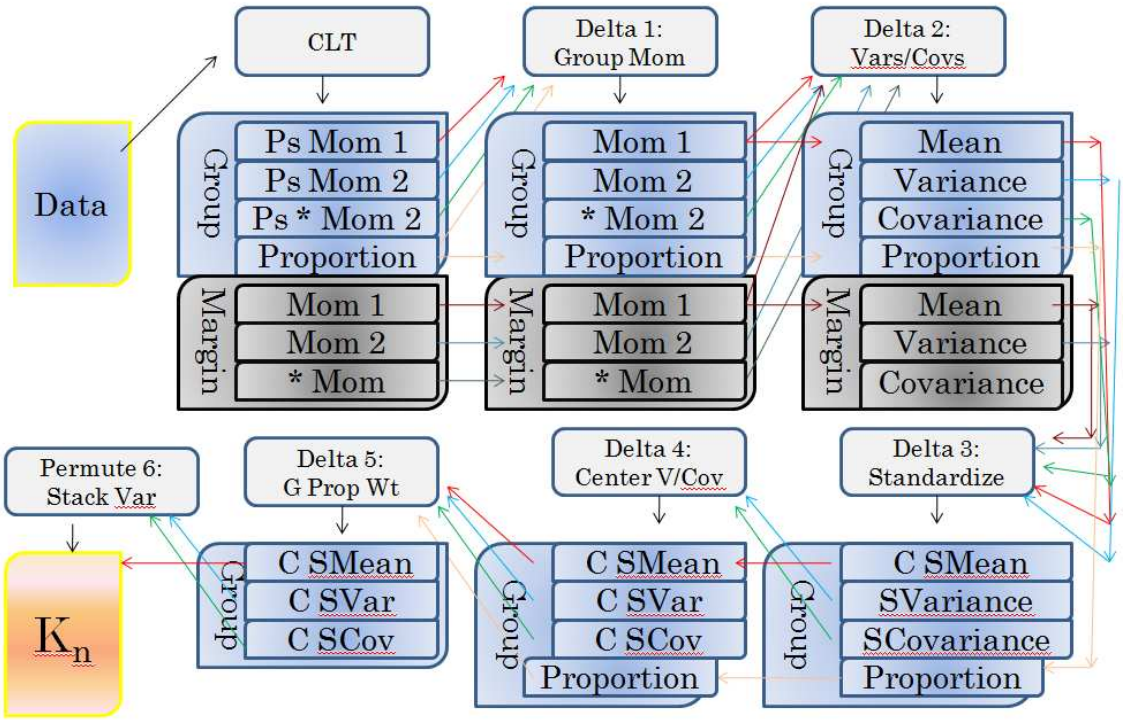

Figure 5. Spanset Schematic

A. Central Limit Theorem

We begin to find the asymptotic distribution of $\mathbf{K_n}$ with a central limit theorem. We will parallel parts of the initial development of the SIR_α kernel distribution in Gannoun and Saracco (2003) for this stage and several of the following stages. Figure 6 summarizes our work.

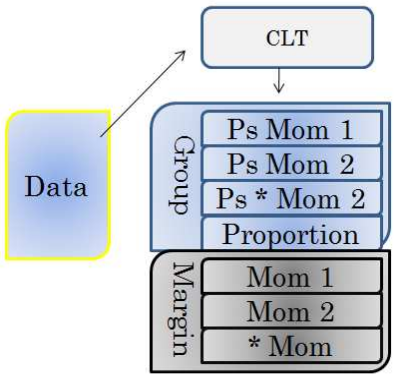

Figure 6. Data To CLT

So we have an i.i.d. sample $(\mathbf{x_1'}, y_1)', \ldots, (\mathbf{x_n'}, y_n)'$ from the population random vectors $(\mathbf{x}', y)'$. The predictors have marginal mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. The response variables, $y_1, \ldots, y_n$ are categorized into g disjoint groups. Each of the groups is a set of real numbers, which inspires the following definition. Define the function $1_i : \mathbb{R} \mapsto \{0, 1\} \ni 1_i(r) = \mathrm{I}(r \in \text{ group } i)$.

For efficiency, we use the shorthand notation $1_{ij} = \mathrm{I}(y_j \in \text{group } i)$. Conditioned on each of the groups, the $\mathbf{x}$ variables are continuous and have finite mean and full rank variance.

For observation i, we define the following $g + gk + gkk + k + kk$ vector.

$$\mathbf{u_i} = \begin{bmatrix} 1_{1i} \\ \vdots \\ 1_{gi} \\ 1_{1i}\mathbf{x_i} \\ \vdots \\ 1_{gi}\mathbf{x_i} \\ 1_{1i}\text{vec}\left(\mathbf{x_i}\mathbf{x_i'}\right) \\ \vdots \\ 1_{gi}\text{vec}\left(\mathbf{x_i}\mathbf{x_i'}\right) \\ \mathbf{x_i} \\ \text{vec}\left(\mathbf{x_i}\mathbf{x_i'}\right) \end{bmatrix} \tag{3.3}$$

We will use the multivariate central limit theorem to ascertain the asymptotic distribution of $\bar{\mathbf{u}}$, which estimates population group proportions, means, and second moments. We will be able to fix the group mean and group second moment estimators later. Now their divisor is the entire sample size n rather than the sample size in their group. The estimated spanning vectors of the SMVCIR space are a function of the elements of $\bar{\mathbf{u}}$. We will use the asymptotic distribution of $\bar{\mathbf{u}}$ to find the asymptotic distribution of vec($\mathbf{K_n}$), and thus $\mathbf{K_n}$.

The expectation of $\mathbf{u_i}$ and therefore $\bar{\mathbf{u}}$ is the following.

$$\mathrm{E}\left[\mathbf{u_i}\right] = \begin{bmatrix} P(y_i \in \text{group1}) \\ \vdots \\ P(y_i \in \text{group}g) \\ \mathrm{E}\left[1_{1i}\mathbf{x_i}\right] \\ \vdots \\ \mathrm{E}\left[1_{gi}\mathbf{x_i}\right] \\ \mathrm{E}\left[1_{1i}\mathrm{vec}\left(\mathbf{x_i}\mathbf{x_i'}\right)\right] \\ \vdots \\ \mathrm{E}\left[1_{gi}\mathrm{vec}\left(\mathbf{x_i}\mathbf{x_i'}\right)\right] \\ \mathrm{E}\left[\mathbf{x_i}\right] \\ \mathrm{E}\left[\mathrm{vec}\left(\mathbf{x_i}\mathbf{x_i'}\right)\right] \end{bmatrix} \tag{3.4}$$

It will help our explanation to create some simplified notation. For the group j, let p_j be $P(y_i \in \text{ group } j)$. This is the proportion of the population that falls into group j.

Let $\mathbf{m_j} = \mathrm{E}\left[1_{ji}\mathbf{x_i}\right] = (m_{j1}, \ldots m_{jk})'$, this is the first moment of the predictors in group j, sans a scaling constant. The nonzero range of the expectation is restricted to population values that fall in group j, but the range of integration of the expectation is all of the groups. In the next chapter we will see how to get the first moment of group j, ignoring the other groups.

$$\mathbf{V_j} = \mathrm{E}\left[1_{ji}\mathbf{x_i}\mathbf{x_i'}\right] = \begin{bmatrix} v_{j11} & \cdots & v_{j1k} \\ \vdots & \ddots & \vdots \\ v_{jk1} & \cdots & v_{jkk} \end{bmatrix} = \begin{bmatrix} v_{j11} & \cdots & v_{jk1} \\ \vdots & \ddots & \vdots \\ v_{j1k} & \cdots & v_{jkk} \end{bmatrix} \tag{3.5}$$

For group j, $\mathbf{V_j}$ gives the second moment of the predictors, with the same scaling constant caveat that applied to $\mathbf{m_j}$.

We will work with $\mathrm{vec}\,(\mathbf{V_j})$, and use the notation of the last matrix in (3.5). The placement of second moment elements in the stacked matrix $\mathrm{E}\,[\mathbf{u_i}]$ is given in ascending order of its group, row index, and finally column index in the second to last matrix in (3.5). Because the second moment matrix is symmetric, this is perfectly accurate. We are simply stacking the rows of the matrix rather than the columns.

The expectation $\mathrm{E}\,[\mathbf{x_i}] = \boldsymbol{\mu}$, the marginal first moment of the predictors and $\mathrm{E}\,[\mathrm{vec}\,(\mathbf{x_i}\mathbf{x_i'})] = \mathrm{vec}\,(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}')$ the marginal second moment. These quantities and their estimates will be used for the standardization.

The variance of $\mathbf{u_i}$ is calculated in Gannoun and Saracco (2003).

$$\mathrm{Var}\,[\mathbf{u_i}] = \begin{bmatrix} \mathbf{B_{pp}} & \mathbf{B_{pm}} & \mathbf{B_{pv}} & \mathbf{B_{pM}} & \mathbf{B_{pV}} \\ \mathbf{B_{mp}} & \mathbf{B_{mm}} & \mathbf{B_{mv}} & \mathbf{B_{mM}} & \mathbf{B_{mV}} \\ \mathbf{B_{vp}} & \mathbf{B_{vm}} & \mathbf{B_{vv}} & \mathbf{B_{vM}} & \mathbf{B_{vV}} \\ \mathbf{B_{Mp}} & \mathbf{B_{Mm}} & \mathbf{B_{Mv}} & \mathbf{B_{MM}} & \mathbf{B_{MV}} \\ \mathbf{B_{Vp}} & \mathbf{B_{Vm}} & \mathbf{B_{Vv}} & \mathbf{B_{VM}} & \mathbf{B_{VV}} \end{bmatrix} \tag{3.6}$$

The matrix $\mathbf{B_{pp}}$ is the variance matrix of the 1_{ji}.

$$\mathbf{B_{pp}} = \begin{bmatrix} p_1(1-p_1) & -p_2p_1 & \cdots & -p_gp_1 \\ -p_1p_2 & p_2(1-p_2) & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ -p_1p_g & \cdots & \cdots & p_g(1-p_g) \end{bmatrix} \tag{3.7}$$

The matrix $\mathbf{B_{pm}}$ is the covariance matrix of the 1_{ji} with the $1_{ji}\mathbf{x_i}$.

$$\mathbf{B_{pm}} = \begin{bmatrix} (1-p_1)\mathbf{m_1'} & -p_1\mathbf{m_2'} & \cdots & -p_1\mathbf{m_g'} \\ -p_2\mathbf{m_1'} & (1-p_2)\mathbf{m_2'} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ -p_g\mathbf{m_1'} & \cdots & \cdots & (1-p_g)\mathbf{m_g'} \end{bmatrix} = \mathbf{B_{mp}'} \tag{3.8}$$

The matrix $\mathbf{B}_{pv}$ is the covariance matrix of the 1_{ji} with the $1_{ji}\text{vec}(\mathbf{x_i x_i'})$.

$$\mathbf{B_{pv}} = \begin{bmatrix} (1-p_1)\text{vec}(\mathbf{V_1})' & -p_1\text{vec}(\mathbf{V_2})' & \cdots & -p_1\text{vec}(\mathbf{V_g})' \\ -p_2\text{vec}(\mathbf{V_1})' & (1-p_2)\text{vec}(\mathbf{V_2})' & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ -p_g\text{vec}(\mathbf{V_1})' & \cdots & \cdots & (1-p_g)\text{vec}(\mathbf{V_g})' \end{bmatrix} = \mathbf{B'_{vp}} \quad (3.9)$$

The matrix $\mathbf{B_{pM}}$ is the covariance matrix of the 1_{ji} with $\mathbf{x_i}$.

$$\mathbf{B_{pM}} = \begin{bmatrix} \mathbf{m_1'} - p_1\boldsymbol{\mu}' \\ \vdots \\ \mathbf{m_g'} - p_g\boldsymbol{\mu}' \end{bmatrix} = \mathbf{B'_{Mp}} \quad (3.10)$$

The matrix $\mathbf{B_{pV}}$ is the covariance matrix of the 1_{ji} with $\text{vec}\,(\mathbf{x_i x_i'})$.

$$\mathbf{B_{pV}} = \begin{bmatrix} \mathbf{vec}\,(\mathbf{V_1})' - p_1\text{vec}\,(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}')' \\ \vdots \\ \mathbf{vec}\,(\mathbf{V_g})' - p_g\text{vec}\,(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}')' \end{bmatrix} = \mathbf{B'_{Vp}} \quad (3.11)$$

The matrix $\mathbf{B_{mm}}$ is the variance matrix of the $1_{ji}\mathbf{x_i}$.

$$\mathbf{B_{mm}} = \begin{bmatrix} \mathbf{V}_1 - \mathbf{m_1 m_1'} & -\mathbf{m_1 m_2'} & \cdots & -\mathbf{m_1 m_g'} \\ -\mathbf{m_2 m_1'} & \mathbf{V}_2 - \mathbf{m_2 m_2'} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ -\mathbf{m_g m_1'} & \cdots & \cdots & \mathbf{V_g} - \mathbf{m_g m_g'} \end{bmatrix} \quad (3.12)$$

Define $\mathbf{M_j} = \text{E}[1_{ji}\mathbf{x_i}(\mathbf{x_i'} \otimes \mathbf{x_i'})]$. The matrix $\mathbf{B_{mv}}$ is the covariance matrix of the

$1_{ji}\mathbf{x_i}$ with the $1_{ji}\text{vec}\,(\mathbf{x_i}\mathbf{x_i}')$.

$$\mathbf{B_{mv}} = \mathbf{B'_{vm}} = \begin{bmatrix} \mathbf{M_1} - \mathbf{m_1}\text{vec}\,(\mathbf{V_1})' & -\mathbf{m_1}\text{vec}\,(\mathbf{V_2})' & \cdots & -\mathbf{m_1}\text{vec}\,(\mathbf{V_g})' \\ -\mathbf{m_2}\text{vec}\,(\mathbf{V_1})' & \mathbf{M_2} - \mathbf{m_2}\text{vec}\,(\mathbf{V_2})' & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ -\mathbf{m_g}\text{vec}\,(\mathbf{V_1})' & \cdots & \cdots & \mathbf{M_g} - \mathbf{m_g}\text{vec}\,(\mathbf{V_g})' \end{bmatrix} \tag{3.13}$$

The matrix $\mathbf{B_{mM}}$ is the covariance matrix of the $1_{ji}\mathbf{x_i}$ with $\mathbf{x_i}$.

$$\mathbf{B_{mM}} = \begin{bmatrix} \mathbf{V}_1' - \mathbf{m_1}\boldsymbol{\mu}' \\ \vdots \\ \mathbf{V_g}' - \mathbf{m_g}\boldsymbol{\mu}' \end{bmatrix} = \mathbf{B_{Mm'}} \tag{3.14}$$

The matrix $\mathbf{B_{mV}}$ is the covariance matrix of the $1_{ji}\mathbf{x_i}$ with $\text{vec}\,(\mathbf{x_i}\mathbf{x_i}')$.

$$\mathbf{B_{mV}} = \begin{bmatrix} \mathbf{M_1} - \mathbf{m_1}\text{vec}\,(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}')' \\ \vdots \\ \mathbf{M_g} - \mathbf{m_g}\text{vec}\,(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}')' \end{bmatrix} = \mathbf{B'_{Vm}} \tag{3.15}$$

Now define $\mathbf{N_j} = \text{E}\left[1_{ji}(\mathbf{x_i}\mathbf{x_i}') \otimes (\mathbf{x_i}\mathbf{x_i}')\right]$. The matrix $\mathbf{B_{vv}}$ is the covariance matrix of the $1_{ji}\text{vec}\,(\mathbf{x_i}\mathbf{x_i}')$.

$$\mathbf{B_{vv}} = \begin{bmatrix} \mathbf{N_1} - \text{vec}(\mathbf{V_1})\text{vec}(\mathbf{V_1})' & -\text{vec}(\mathbf{V_1})\text{vec}(\mathbf{V_2})' & \cdots & -\text{vec}(\mathbf{V_1})\text{vec}(\mathbf{V_g})' \\ -\text{vec}(\mathbf{V_2})\text{vec}(\mathbf{V_1})' & \mathbf{N_2} - \text{vec}(\mathbf{V_2})\text{vec}(\mathbf{V}_2)' & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ -\text{vec}(\mathbf{V_g})\text{vec}(\mathbf{V_1})' & \cdots & \cdots & \mathbf{N_g} - \text{vec}(\mathbf{V_g})\text{vec}\,(\mathbf{V_g})' \end{bmatrix} \tag{3.16}$$

The matrix $\mathbf{B_{Mv}}$ is the covariance matrix of $\mathbf{x_i}$ with the $1_{ji}\text{vec}\,(\mathbf{x_i}\mathbf{x_i}')$.

$$\mathbf{B_{Mv}} = \begin{bmatrix} \mathbf{M_1} - \boldsymbol{\mu}\text{vec}\,(\mathbf{V_1})' & \cdots & \mathbf{M_g} - \boldsymbol{\mu}\text{vec}\,(\mathbf{V_g})' \end{bmatrix} = \mathbf{B'_{vM}} \tag{3.17}$$

The matrix $\mathbf{B_{Vv}}$ is the covariance matrix of $\text{vec}\left(\mathbf{x_i x_i'}\right)$ with the $1_{ji}\text{vec}\left(\mathbf{x_i x_i'}\right)$.

$$\begin{gathered}\mathbf{B_{Vv}} = \mathbf{B_{vV'}} = \\ \left[\begin{array}{ccc} \mathbf{N_1} - \text{vec}\left(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}'\right)\text{vec}\left(\mathbf{V_1}\right)' & \cdots & \mathbf{N_g} - \text{vec}\left(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}'\right)\text{vec}\left(\mathbf{V_g}\right)' \end{array}\right]\end{gathered} \tag{3.18}$$

The matrix $\mathbf{B_{MM}}$ is the variance matrix of $\mathbf{x_i}$. Its value is not surprising.

$$\mathbf{B_{MM}} = \boldsymbol{\Sigma} \tag{3.19}$$

Define $\mathbf{M} = \text{E}\left[\mathbf{x_i}(\mathbf{x_i'} \otimes \mathbf{x_i'})\right]$. The matrix $\mathbf{B_{MV}}$ is the covariance matrix of $\mathbf{x_i}$ with $\text{vec}\left(\mathbf{x_i x_i'}\right)$.

$$\mathbf{B_{MV}} = \mathbf{M} - \boldsymbol{\mu}\text{vec}\left(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}'\right)' \tag{3.20}$$

Define $\mathbf{N} = \text{E}\left[\left(\mathbf{x_i x_i'}\right) \otimes \left(\mathbf{x_i x_i'}\right)\right]$. The matrix $\mathbf{B_{VV}}$ is the variance matrix of $\text{vec}\left(\mathbf{x_i x_i'}\right)$.

$$\mathbf{B_{VV}} = \mathbf{N} - \text{vec}\left(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}'\right)\text{vec}\left(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}'\right)' \tag{3.21}$$

Now we invoke the multivariate central limit theorem. This gives us our first asymptotic distribution.

$$\sqrt{n}\left(\bar{\mathbf{u}} - \text{E}\left[\mathbf{u_i}\right]\right) \rightarrow^d N\left(0, \text{Var}\left[\mathbf{u_i}\right]\right) \tag{3.22}$$

Remember how we are actually only estimating the group moments with a scaling constant. We will fix this in the next chapter, as we get closer to finding the asymptotic distribution of the estimated SMVCIR spanning set. In the following sections, we repeatedly use the delta method. The output from one delta method becomes the input to the next. By the chain rule (as reviewed in the appendix of Lütkepohl (2007)), our repeated invocations are equivalent to one single complex delta method. Explanation is simpler if we treat each stage of the chain separately.

B. Group Moments, Delta Method 1

To fix our estimators so that they are unbiased for the first and second group moments, we need to rescale them. To estimate the mean in group j, we add all the predictor data vectors whose response falls in group j and then divide by their quantity, $n_j^{-1} \sum_1^{n_j} \mathbf{x_i}$. The estimator from the last section, the sample mean of $1_{ji}\mathbf{x_i}$, is actually $n^{-1} \sum_1^{n_j} \mathbf{x_i}$. Figure 7 summarizes our work.

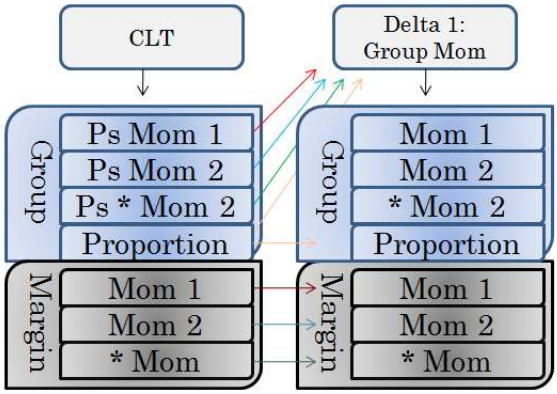

Figure 7. CLT To Group Moments

It is necessary to use this estimator in the initial central limit theorem, because there is no a priori knowledge of the population group proportions. But now that we have invoked the central limit theorem, we may combine our estimators and use the delta method to find the asymptotic distribution of the new statistic.

In this case, we need only multiply the first and second moment estimators by the inverse of their sample group proportions, since $(n_j/n)^{-1} n^{-1} \sum_1^{n_j} \mathbf{x_i} = n_j^{-1} \sum_1^{n_j} \mathbf{x_i}$

.

Let $\mathrm{f}_1 : \mathbb{R}^{g+gk+gkk+k+kk} \to \mathbb{R}^{g+gk+gkk+k+kk}$ such that the following holds.

$$\mathrm{f}_1 \begin{bmatrix} p_1 \\ \vdots \\ p_g \\ m_{11} \\ \vdots \\ m_{1k} \\ \vdots \\ m_{g1} \\ \vdots \\ m_{gk} \\ v_{111} \\ \vdots \\ v_{11k} \\ \vdots \\ v_{1k1} \\ \vdots \\ v_{1kk} \\ \vdots \\ v_{g11} \\ \vdots \\ v_{g1k} \\ \vdots \\ v_{gk1} \\ \vdots \\ v_{gkk} \\ \mu_1 \\ \vdots \\ \mu_k \\ v_{11} \\ \vdots \\ v_{1k} \\ \vdots \\ v_{k1} \\ \vdots \\ v_{kk} \end{bmatrix} = \begin{bmatrix} p_1 \\ \vdots \\ p_g \\ p_1^{-1}m_{11} \\ \vdots \\ p_1^{-1}m_{1k} \\ \vdots \\ p_g^{-1}m_{g1} \\ \vdots \\ p_g^{-1}m_{gk} \\ p_1^{-1}v_{111} \\ \vdots \\ p_1^{-1}v_{11k} \\ \vdots \\ p_1^{-1}v_{1k1} \\ \vdots \\ p_1^{-1}v_{1kk} \\ \vdots \\ p_g^{-1}v_{g11} \\ \vdots \\ p_g^{-1}v_{g1k} \\ \vdots \\ p_g^{-1}v_{gk1} \\ \vdots \\ p_g^{-1}v_{gkk} \\ \mu_1 \\ \vdots \\ \mu_k \\ v_{11} \\ \vdots \\ v_{1k} \\ \vdots \\ v_{k1} \\ \vdots \\ v_{kk} \end{bmatrix} = \begin{bmatrix} p_1 \\ \vdots \\ p_g \\ \mu_{11} \\ \vdots \\ \mu_{1k} \\ \vdots \\ \mu_{g1} \\ \vdots \\ \mu_{gk} \\ \upsilon_{111} \\ \vdots \\ \upsilon_{11k} \\ \vdots \\ \upsilon_{1k1} \\ \vdots \\ \upsilon_{1kk} \\ \vdots \\ \upsilon_{g11} \\ \vdots \\ \upsilon_{g1k} \\ \vdots \\ \upsilon_{gk1} \\ \vdots \\ \upsilon_{gkk} \\ \mu_1 \\ \vdots \\ \mu_k \\ v_{11} \\ \vdots \\ v_{1k} \\ \vdots \\ v_{k1} \\ \vdots \\ v_{kk} \end{bmatrix} \tag{3.23}$$

To invoke the delta method, we must calculate the matrix of derivatives of $\mathrm{f}_1(\mathrm{E}[\mathbf{u_i}])$ with respect to $\mathrm{E}[\mathbf{u_i}]$, $\mathbf{D_1} = \left[\frac{\partial \mathrm{f}_1(\mathrm{E}[\mathbf{u_p}])_i}{\partial \mathrm{E}[\mathbf{u_p}]_j}\right]$. It is necessary that $\mathrm{tr}\left(\mathbf{D_1 D_1'}\right) = \sum_{i=1}^{g+gk+gkk+k+kk} \sum_{j=1}^{g+gk+gkk+k+kk} \left(\mathbf{D_{1ij}}\right)^2 > 0$ for the delta method to work. This condition is clearly met here. Now we will fully define define $\mathbf{D_1}$.

$$\mathbf{D_1} = \begin{bmatrix} \mathbf{I_g} & \mathbf{0} & \mathbf{0} \\ \mathbf{D\boldsymbol{\mu}\boldsymbol{\upsilon}Dp} & \mathbf{D\boldsymbol{\mu}\boldsymbol{\upsilon}Dmv} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I_{k+kk}} \end{bmatrix} \tag{3.24}$$

The matrix $\mathbf{D\boldsymbol{\mu}\boldsymbol{\upsilon}Dp}$ is $gk + gkk \times g$. It contains the derivatives of the group mean and second moment parameters with respect to the group proportions.

$$\mathbf{D\boldsymbol{\mu}\boldsymbol{\upsilon}Dp} = \begin{bmatrix} \mathbf{D\boldsymbol{\mu}_1 Dp} \\ \vdots \\ \mathbf{D\boldsymbol{\mu}_g Dp} \\ \mathbf{D\boldsymbol{\upsilon}_{11} Dp} \\ \vdots \\ \mathbf{D\boldsymbol{\upsilon}_{1k} Dp} \\ \vdots \\ \mathbf{D\boldsymbol{\upsilon}_{g1} Dp} \\ \vdots \\ \mathbf{D\boldsymbol{\upsilon}_{gk} Dp} \end{bmatrix} \tag{3.25}$$

The submatrices $\mathbf{D\boldsymbol{\mu}_j Dp}$ are each $k \times g$.

$$\left[\mathbf{D\boldsymbol{\mu}_l Dp}\right]_{ij} = \frac{\partial \mu_{li}}{\partial p_j} = \frac{\partial \left(m_{li}/p_l\right)}{\partial p_j} = -\frac{m_{li}}{p_j^2}\mathrm{I}(j = l) \tag{3.26}$$

The submatrices $\mathbf{D\boldsymbol{\upsilon}_{li} Dp}$ are also $k \times g$.

$$\left[\mathbf{D\boldsymbol{\upsilon}_{li} Dp}\right]_{fj} = \frac{\partial \upsilon_{lif}}{\partial p_j} = \frac{\partial \left(\upsilon_{lif}/p_l\right)}{\partial p_j} = -\frac{\upsilon_{lif}}{p_j^2}\mathrm{I}(j = l) \tag{3.27}$$

The matrix $\mathbf{D\boldsymbol{\mu\upsilon}Dmv}$ is $gk + gkk \times gk + gkk$. It contains the derivatives of the group means and second moments with respect to the $\mathbf{m_j}$ and $\mathbf{V_j}$ parameters.

$$\mathbf{D\boldsymbol{\mu\upsilon}Dmv} = \text{diag}\begin{bmatrix} \mathbf{D\boldsymbol{\mu}_1 Dm_1} \\ \vdots \\ \mathbf{D\boldsymbol{\mu}_g Dm_g} \\ \mathbf{D\boldsymbol{\upsilon}_{11} Dv_{11}} \\ \vdots \\ \mathbf{D\boldsymbol{\upsilon}_{1k} Dv_{1k}} \\ \vdots \\ \mathbf{D\boldsymbol{\upsilon}_{g1} Dv_{g1}} \\ \vdots \\ \mathbf{D\boldsymbol{\upsilon}_{gk} Dv_{gk}} \end{bmatrix} \tag{3.28}$$

The matrices $\mathbf{D\boldsymbol{\mu}_l Dm_l}$ and $\mathbf{D\boldsymbol{\upsilon}_{lj} Dv_{lj}}$ are each $k \times k$.

$$\mathbf{D\boldsymbol{\mu}_l Dm_l} = \mathbf{D\boldsymbol{\upsilon}_{lj} Dv_{lj}} = \frac{1}{p_l}\mathbf{I_k} \tag{3.29}$$

Since the non-zero derivative condition (mentioned a few pages back) is clearly met, we may conclude the following.

$$\sqrt{n}\left(\mathrm{f}_1\left(\bar{\mathbf{u}}\right) - \mathrm{f}_1\left(\mathrm{E}\left[\mathbf{u_i}\right]\right)\right) \rightarrow^d N\left(0, \mathbf{D_1}\mathrm{Var}\left[\mathbf{u_i}\right]\mathbf{D_1'}\right) \tag{3.30}$$

C. Variances, Delta Method 2

We have first and second group moments estimated in $\mathrm{f}_1\left(\bar{\mathbf{u}}\right)$. We need to turn the second moment estimates into variance estimates. We can do this by combining our estimators and invoking the delta method a second time. Figure 8 summarizes our work.

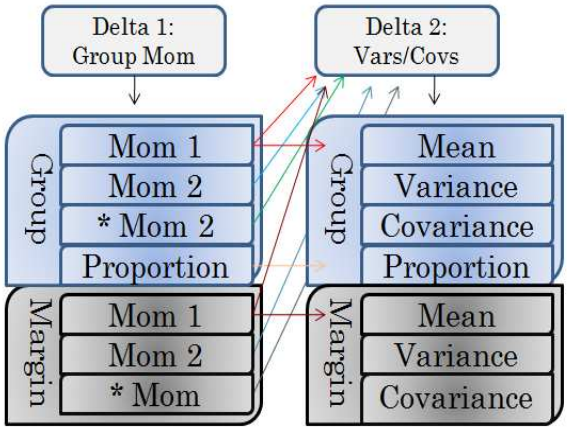

Figure 8. Group Moments to Variances/Covariances

The sample variance can be expanded into an expression that utilizes the second sample moment and sample mean.

$$\begin{aligned}
&\tfrac{1}{n-1}\textstyle\sum_{i=1}^{n}\left(\mathbf{x_i}-\bar{\mathbf{x}}\right)\left(\mathbf{x_i}-\bar{\mathbf{x}}\right)'=\\
&\tfrac{1}{n-1}\textstyle\sum_{i=1}^{n}\left(\mathbf{x_i}\left(\mathbf{x_i}-\bar{\mathbf{x}}\right)'-\bar{\mathbf{x}}\left(\mathbf{x_i}-\bar{\mathbf{x}}\right)'\right)=\\
&\tfrac{1}{n-1}\left(\textstyle\sum_{i=1}^{n}\mathbf{x_i}\mathbf{x_i}'-\left(\sum_{i=1}^{n}\mathbf{x_i}\right)\bar{\mathbf{x}}'-\bar{\mathbf{x}}\left(\sum_{i=1}^{n}\mathbf{x_i}'\right)+n\bar{\mathbf{x}}\bar{\mathbf{x}}'\right)=\\
&\tfrac{1}{n-1}\left(\textstyle\sum_{i=1}^{n}\mathbf{x_i}\mathbf{x_i}'-n\bar{\mathbf{x}}\bar{\mathbf{x}}'-n\bar{\mathbf{x}}\bar{\mathbf{x}}'+n\bar{\mathbf{x}}\bar{\mathbf{x}}'\right)=\\
&\tfrac{1}{n-1}\left(\textstyle\sum_{i=1}^{n}\mathbf{x_i}\mathbf{x_i}'-n\bar{\mathbf{x}}\bar{\mathbf{x}}'\right)
\end{aligned}\tag{3.31}$$

Because of Slutsky's Theorem we have the following.

$$\begin{aligned}
&\tfrac{1}{n-1}\left(\textstyle\sum_{i=1}^{n}\mathbf{x_i}\mathbf{x_i}'-n\bar{\mathbf{x}}\bar{\mathbf{x}}'\right)\approx^{d}\\
&\tfrac{1}{n}\left(\textstyle\sum_{i=1}^{n}\mathbf{x_i}\mathbf{x_i}'-n\bar{\mathbf{x}}\bar{\mathbf{x}}'\right)=\\
&\overline{\mathbf{x}\mathbf{x}'}-\bar{\mathbf{x}}\bar{\mathbf{x}}'
\end{aligned}\tag{3.32}$$

This is a very simple expression in the first and second moment estimators. An expansion of the subtracted $\bar{\mathbf{x}}\bar{\mathbf{x}}'$ term will give us insight into how our transformation function should work for the delta method.

$$\bar{\mathbf{x}}\bar{\mathbf{x}}'=\begin{bmatrix}\bar{x}_1\\ \vdots\\ \bar{x}_k\end{bmatrix}\begin{bmatrix}\bar{x}_1 & \cdots & \bar{x}_k\end{bmatrix}=\begin{bmatrix}\bar{x}_1\bar{x}_1 & \cdots & \bar{x}_k\bar{x}_1\\ \vdots & \ddots & \vdots\\ \bar{x}_1\bar{x}_k & \cdots & \bar{x}_k\bar{x}_k\end{bmatrix}\tag{3.33}$$

Let $\mathrm{f}_2 : \mathbb{R}^{g+gk+gkk+k+kk} \to \mathbb{R}^{g+gk+gkk+k+kk}$ such that

$$\mathrm{f}_2 \begin{bmatrix} p_1 \\ \vdots \\ p_g \\ \mu_{11} \\ \vdots \\ \mu_{1k} \\ \vdots \\ \mu_{g1} \\ \vdots \\ \mu_{gk} \\ \upsilon_{111} \\ \vdots \\ \upsilon_{11k} \\ \vdots \\ \upsilon_{1k1} \\ \vdots \\ \upsilon_{1kk} \\ \vdots \\ \upsilon_{g11} \\ \vdots \\ \upsilon_{g1k} \\ \vdots \\ \upsilon_{gk1} \\ \vdots \\ \upsilon_{gkk} \\ \mu_1 \\ \vdots \\ \mu_k \\ \upsilon_{11} \\ \vdots \\ \upsilon_{1k} \\ \vdots \\ \upsilon_{k1} \\ \vdots \\ \upsilon_{kk} \end{bmatrix} = \begin{bmatrix} p_1 \\ \vdots \\ p_g \\ \mu_{11} \\ \vdots \\ \mu_{1k} \\ \vdots \\ \mu_{g1} \\ \vdots \\ \mu_{gk} \\ \upsilon_{111} - \mu_{11}^2 \\ \vdots \\ \upsilon_{11k} - \mu_{11}\mu_{1k} \\ \vdots \\ \upsilon_{1k1} - \mu_{1k}\mu_{11} \\ \vdots \\ \upsilon_{1kk} - \mu_{1k}^2 \\ \vdots \\ \upsilon_{g11} - \mu_{g1}^2 \\ \vdots \\ \upsilon_{g1k} - \mu_{g1}\mu_{gk} \\ \vdots \\ \upsilon_{gk1} - \mu_{gk}\mu_{g1} \\ \vdots \\ \upsilon_{gkk} - \mu_{gk}^2 \\ \mu_1 \\ \vdots \\ \mu_k \\ \upsilon_{11} - \mu_1^2 \\ \vdots \\ \upsilon_{1k} - \mu_1\mu_k \\ \vdots \\ \upsilon_{k1} - \mu_k\mu_1 \\ \vdots \\ \upsilon_{kk} - \mu_k^2 \end{bmatrix} = \begin{bmatrix} p_1 \\ \vdots \\ p_g \\ \mu_{11} \\ \vdots \\ \mu_{1k} \\ \vdots \\ \mu_{g1} \\ \vdots \\ \mu_{gk} \\ \sigma_{111} \\ \vdots \\ \sigma_{11k} \\ \vdots \\ \sigma_{1k1} \\ \vdots \\ \sigma_{1kk} \\ \vdots \\ \sigma_{g11} \\ \vdots \\ \sigma_{g1k} \\ \vdots \\ \sigma_{gk1} \\ \vdots \\ \sigma_{gkk} \\ \mu_1 \\ \vdots \\ \mu_k \\ \sigma_{11} \\ \vdots \\ \sigma_{1k} \\ \vdots \\ \sigma_{k1} \\ \vdots \\ \sigma_{kk} \end{bmatrix} \tag{3.34}$$

As before, we calculate the matrix of derivatives of $\mathrm{f}_2(\mathrm{f}_1(\mathrm{E}[\mathbf{u_i}]))$ with respect to

$\mathrm{f}_1(\mathrm{E}[\mathbf{u_i}])$, $\mathbf{D_2} = \left[\frac{\partial \mathrm{f}_2(\mathrm{f}_1(\mathrm{E}[\mathbf{u_p}]))_i}{\partial \mathrm{f}_1(\mathrm{E}[\mathbf{u_p}])_j}\right]$. The quantity $\mathrm{tr}(\mathbf{D_2}\mathbf{D_2'})$ is clearly nonzero, so we may invoke the delta method.

$$\mathbf{D_2} = \begin{bmatrix} \mathbf{I_g} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I_{gk}} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{D\sigma D\mu} & \mathbf{I_{gkk}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I_k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{D\sigma m D\mu m} & \mathbf{I_{kk}} \end{bmatrix} \tag{3.35}$$

$$\mathbf{D\sigma D\mu} = \mathrm{diag} \begin{bmatrix} \mathbf{D\sigma_1 D\mu_1} \\ \vdots \\ \mathbf{D\sigma_g D\mu_g} \end{bmatrix} \tag{3.36}$$

$$\mathbf{D\sigma_j D\mu_j} = \begin{bmatrix} \mathbf{D\sigma_{j1} D\mu_j} \\ \vdots \\ \mathbf{D\sigma_{jk} D\mu_j} \end{bmatrix} \tag{3.37}$$

$$\mathbf{D\sigma m D\mu m} = \begin{bmatrix} \mathbf{D\sigma_1 D\mu} \\ \vdots \\ \mathbf{D\sigma_k D\mu} \end{bmatrix} \tag{3.38}$$

The matrices $\mathbf{D\sigma_{li} D\mu_l}$ are $k \times k$.

$$\begin{aligned} &[\mathbf{D\sigma_{li} D\mu_l}]_{fj} = \frac{\partial \sigma_{lif}}{\partial \mu_{lj}} = \frac{\partial\left(v_{lif} - \mu_{li}\mu_{lf}\right)}{\partial \mu_{lj}} = \\ &-\mu_{lf}\mathrm{I}(j = i \neq f) - 2\mu_{li}\mathrm{I}(j = i = f) - \mu_{li}\mathrm{I}(j = f \neq i) \end{aligned} \tag{3.39}$$

The matrices $\mathbf{D}\boldsymbol{\sigma}_\mathbf{i}\mathbf{D}\boldsymbol{\mu}$ are $k \times k$.

$$\begin{aligned}&[\mathbf{D}\boldsymbol{\sigma}_\mathbf{i}\mathbf{D}\boldsymbol{\mu}]_{fj} = \frac{\partial \sigma_{if}}{\partial \mu_j} = \frac{\partial\left(v_{if} - \mu_i\mu_f\right)}{\partial \mu_j} = \\ &-\mu_f \mathrm{I}(j = i \neq f) - 2\mu_i \mathrm{I}(j = i = f) - \mu_i \mathrm{I}(j = f \neq i)\end{aligned} \tag{3.40}$$

Since the non-zero derivative condition is clearly met, we may conclude the following.

$$\sqrt{n}\left(\mathrm{f}_2\left(\mathrm{f}_1\left(\bar{\mathbf{u}}\right)\right) - \mathrm{f}_2\left(\mathrm{f}_1\left(\mathrm{E}\left[\mathbf{u_i}\right]\right)\right)\right) \rightarrow^d N\left(0, \mathbf{D_2}\mathbf{D_1}\mathrm{Var}\left[\mathbf{u_i}\right]\mathbf{D_1'}\mathbf{D_2'}\right) \tag{3.41}$$

So now we have sample group means, variances, and proportions in our spanning set. We need to hold onto the proportions a bit longer, they will be used to weight the spanning set vectors at the end.

D. Standardization, Delta Method 3

Now we need to standardize the data using the scaling transformation. This is accomplished with another delta method. The sample group means will be centered using the marginal sample mean, and then scaled by the inverse of the marginal standard deviations. The sample group variances (covariances) will be scaled by the inverse of the marginal variances (products of standard deviations). Figure 9 summarizes our work.

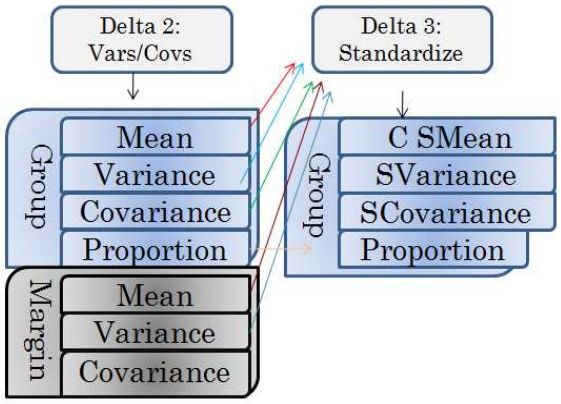

Figure 9. Variances/Covariances to Standardization

Let $\mathrm{f}_2 : \mathbb{R}^{g+gk+gkk+k+kk} \to \mathbb{R}^{g+gk+gkk}$ such that the following holds.

$$\mathrm{f}_3 \begin{bmatrix} p_1 \\ \vdots \\ p_g \\ \mu_{11} \\ \vdots \\ \mu_{1k} \\ \vdots \\ \mu_{g1} \\ \vdots \\ \mu_{gk} \\ \sigma_{111} \\ \vdots \\ \sigma_{11k} \\ \vdots \\ \sigma_{1k1} \\ \vdots \\ \sigma_{1kk} \\ \vdots \\ \sigma_{g11} \\ \vdots \\ \sigma_{g1k} \\ \vdots \\ \sigma_{gk1} \\ \vdots \\ \sigma_{gkk} \\ \mu_1 \\ \vdots \\ \mu_k \\ \sigma_{11} \\ \vdots \\ \sigma_{1k} \\ \vdots \\ \sigma_{k1} \\ \vdots \\ \sigma_{kk} \end{bmatrix} = \begin{bmatrix} p_1 \\ \vdots \\ p_g \\ (\mu_{11} - \mu_1)/\sqrt{\sigma_{11}} \\ \vdots \\ (\mu_{1k} - \mu_k)/\sqrt{\sigma_{kk}} \\ \vdots \\ (\mu_{g1} - \mu_1)/\sqrt{\sigma_{11}} \\ \vdots \\ (\mu_{gk} - \mu_k)/\sqrt{\sigma_{kk}} \\ \sigma_{111}/\sigma_{11} \\ \vdots \\ \sigma_{11k}/\left(\sqrt{\sigma_{11}}\sqrt{\sigma_{kk}}\right) \\ \vdots \\ \sigma_{1k1}/\left(\sqrt{\sigma_{kk}}\sqrt{\sigma_{11}}\right) \\ \vdots \\ \sigma_{1kk}/\sigma_{kk} \\ \vdots \\ \sigma_{g11}/\sigma_{11} \\ \vdots \\ \sigma_{g1k}/\left(\sqrt{\sigma_{11}}\sqrt{\sigma_{kk}}\right) \\ \vdots \\ \sigma_{gk1}/\left(\sqrt{\sigma_{kk}}\sqrt{\sigma_{11}}\right) \\ \vdots \\ \sigma_{gkk}/\sigma_{kk} \end{bmatrix} \tag{3.42}$$

We calculate the matrix of derivatives of $\mathrm{f}_3(\mathrm{f}_2(\mathrm{f}_1(\mathrm{E}[\mathbf{u_i}])))$ with respect to $\mathrm{f}_2(\mathrm{f}_1(\mathrm{E}[\mathbf{u_i}]))$, $\mathbf{D_3} = \left[\frac{\partial \mathrm{f}_3(\mathrm{f}_2(\mathrm{f}_1(\mathrm{E}[\mathbf{u_p}])))_i}{\partial \mathrm{f}_2(\mathrm{f}_1(\mathrm{E}[\mathbf{u_p}]))_j}\right]$. The quantity $\mathrm{tr}\,(\mathbf{D_3 D_3'})$ is clearly nonzero, so we may invoke the delta method.

$$\mathbf{D_3} = \begin{bmatrix} \mathbf{I_g} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{D\mu^z D\mu} & \mathbf{0} & \mathbf{D\mu^z Dm\mu} & \mathbf{D\mu^z Dm\sigma} \\ \mathbf{0} & \mathbf{0} & \mathbf{D\sigma^z D\sigma} & \mathbf{0} & \mathbf{D\sigma^z Dm\sigma} \end{bmatrix} \tag{3.43}$$

$$\mathbf{D\sigma^z D\sigma} = \mathrm{diag} \begin{bmatrix} \mathbf{D\sigma_1^z D\sigma_1} \\ \vdots \\ \mathbf{D\sigma_g^z D\sigma_g} \end{bmatrix} \tag{3.44}$$

$$\mathbf{D\sigma_i^z D\sigma_i} = \mathrm{diag} \begin{bmatrix} \mathbf{D\sigma_{i1}^z D\sigma_{i1}} \\ \vdots \\ \mathbf{D\sigma_{ik}^z D\sigma_{ik}} \end{bmatrix} \tag{3.45}$$

The matrices $\mathbf{D\sigma_{ij}^z D\sigma_{ij}}$ are $k \times k$.

$$\begin{aligned} &\left[\mathbf{D\sigma_{ij}^z D\sigma_{ij}}\right]_{\mathbf{lf}} = \frac{\partial\left(\sigma_{ijl}/\left(\sqrt{\sigma_{jj}}\sqrt{\sigma_{ll}}\right)\right)}{\partial \sigma_{ijf}} = \\ &\mathrm{I}(f = l)\left(\sqrt{\sigma_{jj}}\sqrt{\sigma_{ll}}\right)^{-1} \end{aligned} \tag{3.46}$$

The matrix $\mathbf{D\mu^z D\mu}$ is block diagonal with the $k \times k$ matrix $\left[\mathrm{I}\,(i = j)\,\sqrt{\sigma_{ij}}^{\,-1}\right]$ repeated g times. The matrix $\mathbf{D\mu^z Dm\mu}$ is $gk \times k$ with the $k \times k$ matrix $\left[-\mathrm{I}\,(i = j)\,\sqrt{\sigma_{ij}}^{\,-1}\right]$ repeated g times.

$$\mathbf{D\mu^z Dm\sigma} = \begin{bmatrix} \mathbf{D\mu_1^z Dm\sigma} \\ \vdots \\ \mathbf{D\mu_g^z Dm\sigma} \end{bmatrix} \tag{3.47}$$

$$\mathbf{D}\boldsymbol{\mu}_{\mathbf{i}}^{\mathbf{z}}\mathbf{Dm}\boldsymbol{\sigma} = \begin{bmatrix} \mathbf{D}\boldsymbol{\mu}_{\mathbf{i}}^{\mathbf{z}}\mathbf{Dm}\boldsymbol{\sigma}_{\mathbf{1}} \\ \vdots \\ \mathbf{D}\boldsymbol{\mu}_{\mathbf{i}}^{\mathbf{z}}\mathbf{Dm}\boldsymbol{\sigma}_{\mathbf{k}} \end{bmatrix}' \tag{3.48}$$

The matrices $\mathbf{D}\boldsymbol{\mu}_{\mathbf{i}}^{\mathbf{z}}\mathbf{Dm}\boldsymbol{\sigma}_{\mathbf{j}}$ are each $k \times k$.

$$\begin{aligned} &[\mathbf{D}\boldsymbol{\mu}_{\mathbf{i}}^{\mathbf{z}}\mathbf{Dm}\boldsymbol{\sigma}_{\mathbf{j}}]_{lf} = \tfrac{\partial\left(\left(\mu_{il}-\mu_l\right)/\sqrt{\sigma_{ll}}\right)}{\partial\sigma_{jf}} = \\ &-\mathrm{I}(j=l=f)\tfrac{1}{2}\left(\mu_{il}-\mu_l\right)\sigma_{ll}^{-\frac{3}{2}} \end{aligned} \tag{3.49}$$

$$\mathbf{D}\boldsymbol{\sigma}^{\mathbf{z}}\mathbf{Dm}\boldsymbol{\sigma} = \begin{bmatrix} \mathbf{D}\boldsymbol{\sigma}_{\mathbf{1}}^{\mathbf{z}}\mathbf{Dm}\boldsymbol{\sigma} \\ \vdots \\ \mathbf{D}\boldsymbol{\sigma}_{\mathbf{g}}^{\mathbf{z}}\mathbf{Dm}\boldsymbol{\sigma} \end{bmatrix} \tag{3.50}$$

$$\mathbf{D}\boldsymbol{\sigma}_{\mathbf{i}}^{\mathbf{z}}\mathbf{Dm}\boldsymbol{\sigma} = \begin{bmatrix} \mathbf{D}\boldsymbol{\sigma}_{\mathbf{i1}}^{\mathbf{z}}\mathbf{Dm}\boldsymbol{\sigma} \\ \vdots \\ \mathbf{D}\boldsymbol{\sigma}_{\mathbf{ik}}^{\mathbf{z}}\mathbf{Dm}\boldsymbol{\sigma} \end{bmatrix} \tag{3.51}$$

$$\mathbf{D}\boldsymbol{\sigma}_{\mathbf{ij}}^{\mathbf{z}}\mathbf{Dm}\boldsymbol{\sigma} = \begin{bmatrix} \mathbf{D}\boldsymbol{\sigma}_{\mathbf{ij}}^{\mathbf{z}}\mathbf{Dm}\boldsymbol{\sigma}_{\mathbf{1}} \\ \vdots \\ \mathbf{D}\boldsymbol{\sigma}_{\mathbf{ij}}^{\mathbf{z}}\mathbf{Dm}\boldsymbol{\sigma}_{\mathbf{k}} \end{bmatrix}' \tag{3.52}$$

The matrices $\mathbf{D}\boldsymbol{\sigma}_{\mathbf{ij}}^{\mathbf{z}}\mathbf{Dm}\boldsymbol{\sigma}_{\mathbf{f}}$ are each $k \times k$.

$$\begin{aligned} &\left[\mathbf{D}\boldsymbol{\sigma}_{\mathbf{ij}}^{\mathbf{z}}\mathbf{Dm}\boldsymbol{\sigma}_{\mathbf{f}}\right]_{lm} = \tfrac{\partial\left(\sigma_{ijl}/\sqrt{\sigma_{jj}}\sqrt{\sigma_{ll}}\right)}{\partial\sigma_{fm}} = \\ &\mathrm{I}(f=m)\begin{pmatrix} -\mathrm{I}(f=j\neq l)\frac{\sigma_{ijl}}{2\sigma_{jj}^{3/2}\sqrt{\sigma_{ll}}} \\ -\mathrm{I}(f=l\neq j)\frac{\sigma_{ijl}}{2\sigma_{ll}^{3/2}\sqrt{\sigma_{jj}}} \\ -\mathrm{I}(f=j=l)\frac{\sigma_{ijl}}{\sigma_{jj}^{2}} \end{pmatrix} \end{aligned} \tag{3.53}$$

Other than weighting, the mean components of the spanning set are finished. The variances still have to be centered and then finally weighted. We center the variances by subtracting the weighted-average variance estimate. We will perform this in the next chapter.

Since the non-zero derivative condition is clearly met, we may conclude the following.

$$\sqrt{n}\left(\mathrm{f}_3\left(\mathrm{f}_2\left(\mathrm{f}_1\left(\bar{\mathbf{u}}\right)\right)\right)-\mathrm{f}_3\left(\mathrm{f}_2\left(\mathrm{f}_1\left(\mathrm{E}\left[\mathbf{u_i}\right]\right)\right)\right)\right)\rightarrow^{d} N\left(0,\mathbf{D_3D_2D_1}\mathrm{Var}\left[\mathbf{u_i}\right]\mathbf{D_1'D_2'D_3'}\right) \quad (3.54)$$

E. Centered Variance, Delta Method 4

We define the weighted average variance (pooled variance or expected conditional variance) as $\sum_{i=1}^{g} p_i \mathbf{\Sigma_i}$. So the weighted average covariance of the standardized predictor components i and j is $\sum_{l=1}^{g} p_l \sigma_{lij}$. We can use the proportion and variance/covariance components of our estimator to create new variance/covariance components $\breve{\sigma}_{pij}$ that have been centered, $\sigma_{pij} - \sum_{l=1}^{g} p_l \sigma_{lij}$. The distribution of our estimator, with these new variance/covariance components replacing the old ones, is easily determined with another delta method invocation. Figure 10 summarizes our work.

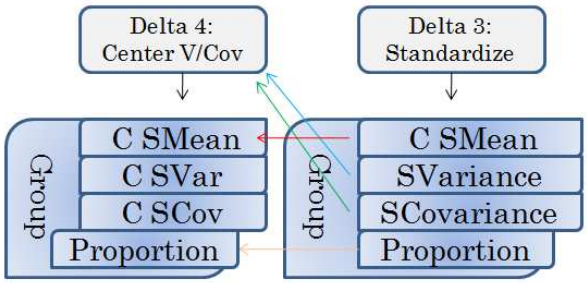

Figure 10. Standardization To Centered Variances/Covariances

Let $\mathrm{f}_4 : \mathbb{R}^{g+gk+gkk} \to \mathbb{R}^{g+gk+gkk}$ such that the following holds.

$$\mathrm{f}_4 \begin{bmatrix} p_1 \\ \vdots \\ p_g \\ \mu_{11} \\ \vdots \\ \mu_{1k} \\ \vdots \\ \mu_{g1} \\ \vdots \\ \mu_{gk} \\ \sigma_{111} \\ \vdots \\ \sigma_{11k} \\ \vdots \\ \sigma_{1k1} \\ \vdots \\ \sigma_{1kk} \\ \vdots \\ \sigma_{g11} \\ \vdots \\ \sigma_{g1k} \\ \vdots \\ \sigma_{gk1} \\ \vdots \\ \sigma_{gkk} \end{bmatrix} = \begin{bmatrix} p_1 \\ \vdots \\ p_g \\ \mu_{11} \\ \vdots \\ \mu_{1k} \\ \vdots \\ \mu_{g1} \\ \vdots \\ \mu_{gk} \\ \sigma_{111} - \sum_{i=1}^{g} p_i \sigma_{i11} = \breve{\sigma}_{111} \\ \vdots \\ \sigma_{11k} - \sum_{i=1}^{g} p_i \sigma_{i1k} = \breve{\sigma}_{11k} \\ \vdots \\ \sigma_{1k1} - \sum_{i=1}^{g} p_i \sigma_{ik1} = \breve{\sigma}_{1k1} \\ \vdots \\ \sigma_{1kk} - \sum_{i=1}^{g} p_i \sigma_{ikk} = \breve{\sigma}_{1kk} \\ \vdots \\ \sigma_{g11} - \sum_{i=1}^{g} p_i \sigma_{i11} = \breve{\sigma}_{g11} \\ \vdots \\ \sigma_{g1k} - \sum_{i=1}^{g} p_i \sigma_{i1k} = \breve{\sigma}_{g1k} \\ \vdots \\ \sigma_{gk1} - \sum_{i=1}^{g} p_i \sigma_{ik1} = \breve{\sigma}_{gk1} \\ \vdots \\ \sigma_{gkk} - \sum_{i=1}^{g} p_i \sigma_{ikk} = \breve{\sigma}_{gkk} \end{bmatrix} \tag{3.55}$$

We calculate the matrix of derivatives of $\mathrm{f}_4(\mathrm{f}_3(\mathrm{f}_2(\mathrm{f}_1(\mathrm{E}[\mathbf{u_i}]))))$ with respect to $\mathrm{f}_3(\mathrm{f}_2(\mathrm{f}_1(\mathrm{E}[\mathbf{u_i}])))$,$\mathbf{D}_4 = \left[\frac{\partial \mathrm{f}_4(\mathrm{f}_3(\mathrm{f}_2(\mathrm{f}_1(\mathrm{E}[\mathbf{u_p}]))))_i}{\partial \mathrm{f}_3(\mathrm{f}_2(\mathrm{f}_1(\mathrm{E}[\mathbf{u_p}])))_j}\right]$. The quantity $\mathrm{tr}\,(\mathbf{D_4 D_4'})$ is clearly nonzero, so we may invoke the delta method.

$$\mathbf{D_4} = \begin{bmatrix} \mathbf{I_g} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I_{gk}} & \mathbf{0} \\ \mathbf{D\breve{\sigma} D p} & \mathbf{0} & \mathbf{D\breve{\sigma} D \sigma} \end{bmatrix} \tag{3.56}$$

The matrix $\mathbf{D\breve{\sigma}Dp}$ contains the derivative of the centered variances with respect to the group proportions.

$$\mathbf{D\breve{\sigma}Dp} = \begin{bmatrix} \mathbf{D\breve{\sigma}_{11}Dp} \\ \vdots \\ \mathbf{D\breve{\sigma}_{1k}Dp} \\ \vdots \\ \mathbf{D\breve{\sigma}_{g1}Dp} \\ \vdots \\ \mathbf{D\breve{\sigma}_{gk}Dp} \end{bmatrix} \tag{3.57}$$

The matrices $\mathbf{D\breve{\sigma}_{1q}Dp}$ are $k \times g$.

$$\left[\mathbf{D\breve{\sigma}_{lq}Dp}\right]_{fj} = \frac{\partial \breve{\sigma}_{lqf}}{\partial p_j} = \frac{\partial \left(\sigma_{lqf} - \sum_{i=1}^{g} p_i \sigma_{iqf}\right)}{\partial p_j} = -\sigma_{jqf} \tag{3.58}$$

The matrix $\mathbf{D\breve{\sigma}D\sigma}$ contains the derivatives of the centered variances with respect to the original variances.

$$\mathbf{D\breve{\sigma}D\sigma} = \begin{bmatrix} \mathbf{D\breve{\sigma}_{11}D\sigma} \\ \vdots \\ \mathbf{D\breve{\sigma}_{1k}D\sigma} \\ \vdots \\ \mathbf{D\breve{\sigma}_{g1}D\sigma} \\ \vdots \\ \mathbf{D\breve{\sigma}_{gk}D\sigma} \end{bmatrix} \tag{3.59}$$

$$\mathbf{D\breve{\sigma}_{1j}D\sigma} = \begin{bmatrix} \mathbf{D\breve{\sigma}_{1j}D\sigma_{11}} & \cdots & \mathbf{D\breve{\sigma}_{1j}D\sigma_{1k}} & \cdots & \mathbf{D\breve{\sigma}_{1j}D\sigma_{g1}} & \cdots & \mathbf{D\breve{\sigma}_{1j}D\sigma_{gk}} \end{bmatrix} \tag{3.60}$$

The matrices $\mathbf{D\breve{\sigma}_{1q}D\sigma_{hm}}$ are $k \times k$

$$\begin{aligned}&[\mathbf{D\breve{\sigma}_{1q}D\sigma_{hm}}]_{fj} = \frac{\partial \breve{\sigma}_{lqf}}{\partial \sigma_{hmj}} = \frac{\partial\left(\sigma_{lqf} - \sum_{i=1}^{g} p_i \sigma_{iqf}\right)}{\partial \sigma_{hmj}} = \\ &(1 - p_l)\,\mathrm{I}(h = l, m = q, j = f) - p_h \mathrm{I}(h \neq l, m = q, j = f)\end{aligned} \tag{3.61}$$

Since the non-zero derivative condition is clearly met, we may conclude the following.

$$\begin{aligned}&\sqrt{n}\left(\mathrm{f}_4(\mathrm{f}_3\left(\mathrm{f}_2\left(\mathrm{f}_1\left(\bar{\mathbf{u}}\right)\right)\right)\right) - \mathrm{f}_4(\mathrm{f}_3\left(\mathrm{f}_2\left(\mathrm{f}_1\left(\mathrm{E}\left[\mathbf{u_i}\right]\right)\right)\right)\right) \rightarrow^d \\ &N\left(0, \mathbf{D_4 D_3 D_2 D_1} \mathrm{Var}\left[\mathbf{u_i}\right] \mathbf{D_1' D_2' D_3' D_4'}\right)\end{aligned} \tag{3.62}$$

F. Weighting, Delta Method 5

We weight each of the statistics by the square root of the group proportion to finalize the estimates of the spanning set components. This can be achieved by multiplying the means and centered variance/covariance estimates by the group proportion estimates. We can use the delta method again. This invocation is similar to the first delta method invocation, save that we are not inverting the proportions. Figure 11 summarizes our work.

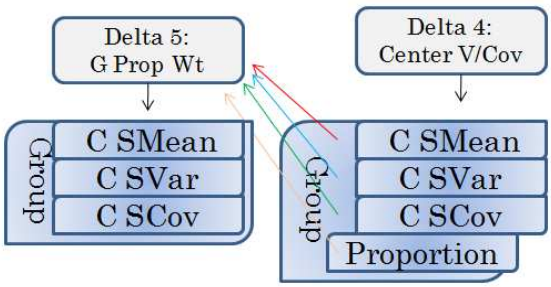

Figure 11. Centered Variances/Covariances To Weighting

Let $\mathrm{f}_4 : \mathbb{R}^{g+gk+gkk} \to \mathbb{R}^{gk+gkk}$ such that the following holds.

$$\mathrm{f}_5 \begin{bmatrix} p_1 \\ \vdots \\ p_g \\ \mu_{11} \\ \vdots \\ \mu_{1k} \\ \vdots \\ \mu_{g1} \\ \vdots \\ \mu_{gk} \\ \breve{\sigma}_{111} \\ \vdots \\ \breve{\sigma}_{11k} \\ \vdots \\ \breve{\sigma}_{1k1} \\ \vdots \\ \breve{\sigma}_{1kk} \\ \vdots \\ \breve{\sigma}_{g11} \\ \vdots \\ \breve{\sigma}_{g1k} \\ \vdots \\ \breve{\sigma}_{gk1} \\ \vdots \\ \breve{\sigma}_{gkk} \end{bmatrix} = \begin{bmatrix} \sqrt{p_1}\mu_{11} \\ \vdots \\ \sqrt{p_1}\mu_{1k} \\ \vdots \\ \sqrt{p_g}\mu_{g1} \\ \vdots \\ \sqrt{p_g}\mu_{gk} \\ \sqrt{p_1}\breve{\sigma}_{111} \\ \vdots \\ \sqrt{p_1}\breve{\sigma}_{11k} \\ \vdots \\ \sqrt{p_1}\breve{\sigma}_{1k1} \\ \vdots \\ \sqrt{p_1}\breve{\sigma}_{1kk} \\ \vdots \\ \sqrt{p_g}\breve{\sigma}_{g11} \\ \vdots \\ \sqrt{p_g}\breve{\sigma}_{g1k} \\ \vdots \\ \sqrt{p_g}\breve{\sigma}_{gk1} \\ \vdots \\ \sqrt{p_g}\breve{\sigma}_{gkk} \end{bmatrix} = \begin{bmatrix} V_{11} \\ \vdots \\ V_{1k} \\ \vdots \\ V_{g1} \\ \vdots \\ V_{gk} \\ \Delta_{111} \\ \vdots \\ \Delta_{11k} \\ \vdots \\ \Delta_{1k1} \\ \vdots \\ \Delta_{1kk} \\ \vdots \\ \Delta_{g11} \\ \vdots \\ \Delta_{g1k} \\ \vdots \\ \Delta_{gk1} \\ \vdots \\ \Delta_{gkk} \end{bmatrix} \tag{3.63}$$

We calculate the matrix of derivatives of $\mathrm{f}_5\left(\mathrm{f}_4\left(\mathrm{f}_3(\mathrm{f}_2(\mathrm{f}_1(\mathrm{E}[\mathbf{u_i}])))\right)\right)$ with respect to $\mathrm{f}_4\left(\mathrm{f}_3\left(\mathrm{f}_2(\mathrm{f}_1(\mathrm{E}[\mathbf{u_i}]))\right)\right)$, $\mathbf{D_5} = \left[\frac{\partial \mathrm{f}_5(\mathrm{f}_4(\mathrm{f}_3(\mathrm{f}_2(\mathrm{f}_1(\mathrm{E}[\mathbf{u_p}])))))_i}{\partial \mathrm{f}_4(\mathrm{f}_3(\mathrm{f}_2(\mathrm{f}_1(\mathrm{E}[\mathbf{u_p}]))))_j}\right]$. The quantity $\mathrm{tr}(\mathbf{D_5}\mathbf{D_5'})$ is clearly nonzero, so we may invoke the delta method.

$$\mathbf{D_5} = \begin{bmatrix} \mathbf{DVDp} & \mathbf{DVD}\boldsymbol{\mu} & \mathbf{0} \\ \mathbf{D\Delta Dp} & \mathbf{0} & \mathbf{D\Delta D}\breve{\boldsymbol{\sigma}} \end{bmatrix} \tag{3.64}$$

The matrix $\mathbf{DVDp}$ contains the derivatives of the weighted mean difference

parameters with respect to the group proportions.

$$\mathbf{DVDp} = \text{diagonal}\begin{bmatrix} \mathbf{DV_1Dp_1} \\ \vdots \\ \mathbf{DV_gDp_g} \end{bmatrix} \tag{3.65}$$

The $\mathbf{DV_jDp_j}$ matrices are $k \times 1$ vectors.

$$\left[\mathbf{DV_jDp_j}\right]_m = \frac{\partial V_{jm}}{\partial p_j} = \frac{\partial\left(\sqrt{p_j}\mu_{jm}\right)}{\partial p_j} = \frac{\mu_{jm}}{2\sqrt{p_j}} \tag{3.66}$$

The matrix $\mathbf{D\Delta Dp}$ holds the derivatives of the weighted variance parameters with respect to the group proportions.

$$\mathbf{D\Delta Dp} = \text{diag}\begin{bmatrix} \mathbf{D\Delta_1Dp_1} \\ \vdots \\ \mathbf{D\Delta_gDp_g} \end{bmatrix} \tag{3.67}$$

$$\mathbf{D\Delta_jDp_j} = \begin{bmatrix} \mathbf{D\Delta_{j1}Dp_j} \\ \vdots \\ \mathbf{D\Delta_{jk}Dp_j} \end{bmatrix} \tag{3.68}$$

The matrices $\mathbf{D\Delta_{ij}Dp_i}$ are $k \times 1$ vectors.

$$\left[\mathbf{D\Delta_{ij}Dp_i}\right]_m = \frac{\partial \Delta_{ijm}}{\partial p_i} = \frac{\partial\left(\sqrt{p_i}\breve{\sigma}_{ijm}\right)}{\partial p_i} = \frac{\breve{\sigma}_{ijm}}{2\sqrt{p_i}} \tag{3.69}$$

The matrix $\mathbf{DVD}\boldsymbol{\mu}$ holds the derivatives of the weighted means with respect to the unweighted means.

$$\mathbf{DVD}\boldsymbol{\mu} = \begin{bmatrix} \sqrt{p_1}\mathbf{I_k} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sqrt{p_g}\mathbf{I_k} \end{bmatrix} \tag{3.70}$$

The matrix $\mathbf{D\Delta D\breve{\sigma}}$ holds the derivatives of the weighted variances with respect to the unweighted variances.

$$\mathbf{D\Delta D\breve{\sigma}} = \begin{bmatrix} \sqrt{p_1}\mathbf{I_{kk}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sqrt{p_g}\mathbf{I_{kk}} \end{bmatrix} \tag{3.71}$$

Since the non-zero derivative condition is clearly met, we may conclude the following.

$$\begin{aligned} &\sqrt{n}\left(\mathrm{f}_5\left(\mathrm{f}_4\left(\mathrm{f}_3\left(\mathrm{f}_2\left(\mathrm{f}_1\left(\bar{\mathbf{u}}\right)\right)\right)\right)\right) - \mathrm{f}_5\left(\mathrm{f}_4\left(\mathrm{f}_3\left(\mathrm{f}_2\left(\mathrm{f}_1\left(\mathrm{E}\left[\mathbf{u_i}\right]\right)\right)\right)\right)\right)\right) \rightarrow^d \\ &N\left(0, \mathbf{D_5 D_4 D_3 D_2 D_1} \mathrm{Var}\left[\mathbf{u_i}\right] \mathbf{D_1' D_2' D_3' D_4' D_5'}\right) \end{aligned} \tag{3.72}$$

We now have estimates of all of the component elements of our spanning set and their asymptotic distribution. We need to reorder the elements and add some zero cells to properly compose the spanning set vectors. Recall that SMVCIR stacks all the variance differences together, separately from the covariance differences.

G. Final Spanset Stage: Permutations & Slutsky's

We will adopt some new notation for simplicity.

$$\begin{aligned} \mathbf{M} &= \mathrm{f}_5\left(\mathrm{f}_4\left(\mathrm{f}_3\left(\mathrm{f}_2\left(\mathrm{f}_1\left(\mathrm{E}\left[\mathbf{u_i}\right]\right)\right)\right)\right)\right) \\ \hat{\mathbf{M}} &= \mathrm{f}_5\left(\mathrm{f}_4\left(\mathrm{f}_3\left(\mathrm{f}_2\left(\mathrm{f}_1\left(\bar{\mathbf{u}}\right)\right)\right)\right)\right) \\ \mathbf{V} &= \mathbf{D_5 D_4 D_3 D_2 D_1} \mathrm{Var}\left[\mathbf{u_i}\right] \mathbf{D_1' D_2' D_3' D_4' D_5'} \end{aligned} \tag{3.73}$$

We want to transform $\mathbf{M}$ and $\hat{\mathbf{M}}$ so that their variance components are stacked together, with 0's in the variance components previous positions. Figure 12 summarizes our work.

First, we add zero cells. This is done via a matrix multiplication. Look at the

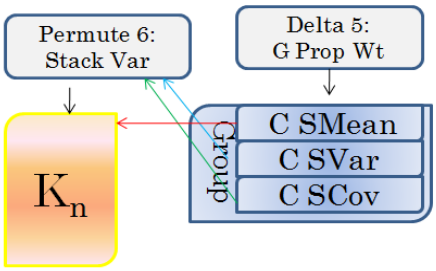

Figure 12. Weighting To Permutation

following example.

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 0 \end{bmatrix} \tag{3.74}$$

This can be easily extended by adding more zero rows on the permutation matrix. If we multiply our estimator by such a zero row augmented permutation matrix, we may invoke Slutsky's Theorem to find its new asymptotic distribution. It will clearly be normal, supposing that zero-variance components are allowed in a multivariate normal random vector. This supposition should not cause difficulties here, so we will make it.

$$\sqrt{n}\left(\begin{bmatrix} \hat{\mathbf{M}} \\ \mathbf{0_{gk}} \end{bmatrix} - \begin{bmatrix} \mathbf{M} \\ \mathbf{0_{gk}} \end{bmatrix}\right) \rightarrow^{d} N\left(\mathbf{0}, \begin{bmatrix} \mathbf{V} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\right) \tag{3.75}$$

Now we permute. We will use some Matlab notation here, as in Golub and Loan (1996). Let $\text{Row}(z) = \mathbf{I_{gk+gkk+gk}}(z, :)$. This function gives the row vector of $\mathbf{I_{gk+gkk+gk}}$ corresponding to the input index.

We will refer to the rows of our $gk + gkk + gk$ permutation matrix, $\mathbf{P}$ using a parenthetic index. So, $\mathbf{P}(i)$ refers to row i of $\mathbf{P}$. We will define $\mathbf{P}$ after some explanatory discussion about the indices of our variance estimates.

Suppose we had $k = 3$. Our variance differences for group i are stacked as so.

$$\begin{array}{cc} 1 & \left[\begin{array}{c} \Delta_{i11} \end{array}\right. \\ 2 & \left|\begin{array}{c} \Delta_{i12} \end{array}\right. \\ 3 & \left|\begin{array}{c} \Delta_{i13} \end{array}\right. \\ 4 & \left|\begin{array}{c} \Delta_{i21} \end{array}\right. \\ 5 & \left|\begin{array}{c} \Delta_{i22} \end{array}\right. \\ 6 & \left|\begin{array}{c} \Delta_{i23} \end{array}\right. \\ 7 & \left|\begin{array}{c} \Delta_{i31} \end{array}\right. \\ 8 & \left|\begin{array}{c} \Delta_{i32} \end{array}\right. \\ 9 & \left.\begin{array}{c} \Delta_{i33} \end{array}\right] \end{array} \tag{3.76}$$

The first variance component is found at index $1 + 0(3 + 1) = 1$.

$$\begin{array}{cc} \mathbf{1} & \left[\begin{array}{c} \mathbf{\Delta_{i11}} \end{array}\right. \\ 2 & \left|\begin{array}{c} \Delta_{i12} \end{array}\right. \\ 3 & \left|\begin{array}{c} \Delta_{i13} \end{array}\right. \\ 4 & \left|\begin{array}{c} \Delta_{i21} \end{array}\right. \\ 5 & \left|\begin{array}{c} \Delta_{i22} \end{array}\right. \\ 6 & \left|\begin{array}{c} \Delta_{i23} \end{array}\right. \\ 7 & \left|\begin{array}{c} \Delta_{i31} \end{array}\right. \\ 8 & \left|\begin{array}{c} \Delta_{i32} \end{array}\right. \\ 9 & \left.\begin{array}{c} \Delta_{i33} \end{array}\right] \end{array} \tag{3.77}$$

The second variance component is found at index $1 + 1(3 + 1) = 5$.

$$\begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ \mathbf{5} \\ 6 \\ 7 \\ 8 \\ 9 \end{matrix} \begin{bmatrix} \Delta_{i11} \\ \Delta_{i12} \\ \Delta_{i13} \\ \Delta_{i21} \\ \mathbf{\Delta_{i22}} \\ \Delta_{i23} \\ \Delta_{i31} \\ \Delta_{i32} \\ \Delta_{i33} \end{bmatrix} \tag{3.78}$$

The third and last variance component is found at index $1 + 2(3 + 1) = 9$.

$$\begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ \mathbf{9} \end{matrix} \begin{bmatrix} \Delta_{i11} \\ \Delta_{i12} \\ \Delta_{i13} \\ \Delta_{i21} \\ \Delta_{i22} \\ \Delta_{i23} \\ \Delta_{i31} \\ \Delta_{i32} \\ \mathbf{\Delta_{i33}} \end{bmatrix} \tag{3.79}$$

This example can be easily extended for higher values of k. Let l_i be the last index before the variance/covariance estimates for group i. The variance estimates for group i are found at indices $l_i + 1 + (z - 1)(k + 1)$ where $z = 1, \ldots, k$.

We can easily calculate these l_i indices upon realizing that the mean differences

account for the first gk differences and that each group's variance/covariance differences are spread over kk indices. So $l_i = gk + kk(i-1)$.

We will switch the variance estimates with the zero cells at the bottom of $\hat{\mathbf{M}}$ by replacing $\hat{\mathbf{M}}$ with $\mathbf{P}\hat{\mathbf{M}}$.

Initially, we let $\mathbf{P_i} = \text{Row}(i)$ for all indices. The mean and covariance estimates will not switch positions from $\hat{\mathbf{M}}$ to $\mathbf{P}\hat{\mathbf{M}}$, and this action fixes them in the correct position after multiplication. The variances are switched with the zero cells using the following formulas.

$$\begin{aligned}
&\mathbf{P}(gk + kk(i-1) + 1 + (z-1)(k+1)) = \text{Row}(gk + gkk + z + (i-1)k) \\
&\mathbf{P}(gk + gkk + z + (i-1)k) = \text{Row}(gk + kk(i-1) + 1 + (z-1)(k+1)) \\
&i = 1, \ldots, g \\
&z = 1, \ldots, k
\end{aligned} \tag{3.80}$$

So by Slutsky's Theorem we obtain the following.

$$\sqrt{n}\left(\mathbf{P}\begin{bmatrix}\hat{\mathbf{M}} \\ \mathbf{0_{gk}}\end{bmatrix} - \mathbf{P}\begin{bmatrix}\mathbf{M} \\ \mathbf{0_{gk}}\end{bmatrix}\right) \rightarrow^d N\left(\mathbf{0}, \mathbf{P}\begin{bmatrix}\mathbf{V} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}\mathbf{P}'\right) \tag{3.81}$$

Now we redefine our notation.

$$\begin{aligned}
\left(\mathbf{P}\begin{bmatrix}\hat{\mathbf{M}} \\ \mathbf{0_{gk}}\end{bmatrix}\right) &= \text{vec}\,(\mathbf{K_n}) \\
\left(\mathbf{P}\begin{bmatrix}\mathbf{M} \\ \mathbf{0_{gk}}\end{bmatrix}\right) &= \text{vec}\,(\mathbf{K}) \\
\left(\mathbf{P}\begin{bmatrix}\mathbf{V} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}\mathbf{P}'\right) &= \mathbf{\Delta}
\end{aligned} \tag{3.82}$$

$$\text{vec}\sqrt{n}\,(\mathbf{K_n} - \mathbf{K}) \rightarrow^d N\,(\mathbf{0}, \mathbf{\Delta})$$

The kernel of the SMVCIR space is then estimated by the following.

$$\mathbf{K_n K_n'} = [\mathbf{k_{n1}} \dots \mathbf{k_{nh}}] \begin{bmatrix} \mathbf{k_{n1}} \\ \vdots \\ \mathbf{k_{nh}} \end{bmatrix} = \mathbf{k_{n1} k_{n1}'} + \cdots + \mathbf{k_{nh} k_{nh}'} \tag{3.83}$$

It is time to show how to make inference on d, the dimension of the SMVCIR space.

CHAPTER IV

DIMENSIONALITY TEST : FINAL TEST STATISTIC

From the last chapter, we have decided theoretically the h vector spanning set $\mathbf{K}$ and the kernel of the SMVCIR space $\mathbf{KK'}$. We also have an estimator for $\mathbf{K}$, $\mathbf{K_n}$ with a known asymptotic distribution.

We will provide a test for the dimensionality of $\mathbf{K}$, d in this chapter.

For our tests of dimensionality of the SMVCIR space, d, we will use both the left and right singular vectors of $\mathbf{K_n}$. Our approach was inspired by that of Yin (2005). The same final statistic and a similar testing scheme is used in both SAVE (Cook and Yin (2001)) and SIR (Li (1991)).

The dimension of the SMVCIR space is the same as the rank of the spanning set matrix. This is the same as the number of non-zero singular values of the spanning set $\mathbf{K}$ and non-zero eigenvalues of the kernel $\mathbf{KK'}$. We saw how to diagonalize $\mathbf{K}$ in (1.19) and $\mathbf{KK'}$ in (1.17).

We suggest a one-sided alternative test. The eigen and singular values are always non-negative, so there is no obvious point of symmetry to base a two sided hypothesis test around.

To test the null hypothesis $d = i$ versus the alternative $d > i$, we could sum the $d + 1, \ldots, k$ diagonal elements of one of the diagonalized estimate matrices just mentioned, and reject the null if the resultant value is too large. Under the null hypothesis, the parameters that these diagonal elements estimate will be zero.

As we discussed before, To actually estimate a value of d we use an iterative

series of tests.

$$\begin{aligned}
&\text{Test } 0: \mathrm{H}_0: d = 0 \text{ vs } \mathrm{H}_1: d > 0 \\
&\text{Test } 1: \mathrm{H}_0: d = 1 \text{ vs } \mathrm{H}_1: d > 1 \\
&\vdots \\
&\text{Test } i: \mathrm{H}_0: d = i \text{ vs } \mathrm{H}_1: d > i \\
&\vdots \\
&\text{Test } k-1: \mathrm{H}_0: d = k-1 \text{ vs } \mathrm{H}_1: d = k
\end{aligned} \tag{4.1}$$

For test i , we calculate the following statistic.

$$\hat{\Lambda}_i = n \sum_{j=i+1}^{k} \hat{\lambda}_j \tag{4.2}$$

The scalar values $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_k$ are the eigenvalues of the estimated kernel matrix $\mathbf{K_n K'_n}$.

We proceed by first writing $\hat{\Lambda}_i$ in terms of the singular value decomposition of the spanning set. Then we find the null distribution of a similar expression that replaces the estimated spanning set statistic with the mean adjusted estimated spanning set statistic from (3.82). Then we show how the distribution of our test statistic and the mean-adjusted statistic are equivalent. Finally we provide 3 reference distributions for the null distribution of our statistic based on these results.

A. Test Statistic $\hat{\Lambda}_i$ in Terms of SVD of Span Set

To perform test i, we have to determine the distribution of $\hat{\Lambda}_i$ under $\mathrm{H}_0 : d = i$. To do this, we first return to the singular value decomposition of $\mathbf{K}$.

$$\begin{aligned}
\mathbf{K} &= \begin{bmatrix} \mathbf{U_1} & \mathbf{U_0} \end{bmatrix} \begin{bmatrix} \mathbf{D_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{D_0} \end{bmatrix} \begin{bmatrix} \mathbf{V_1} & \mathbf{V_0} \end{bmatrix}' \\
\mathbf{U_1} &= \begin{bmatrix} \mathbf{u_1} & \cdots & \mathbf{u_i} \end{bmatrix} \\
\mathbf{U_0} &= \begin{bmatrix} \mathbf{u_{i+1}} & \cdots & \mathbf{u_k} \end{bmatrix} \\
\mathbf{V_1} &= \begin{bmatrix} \mathbf{v_1} & \cdots & \mathbf{v_i} \end{bmatrix} \\
\mathbf{V_0} &= \begin{bmatrix} \mathbf{v_{i+1}} & \cdots & \mathbf{v_h} \end{bmatrix}
\end{aligned} \tag{4.3}$$

The matrix $\mathbf{D_1}$ is diagonal and contains $\sigma_1 \geq \ldots \geq \sigma_i$ on its diagonal. The rectangular matrix $\mathbf{D_0}$ has all zero entries, except for along its diagonal (where the row and column indices are equal), which contains $\sigma_{i+1} \geq \ldots \geq \sigma_k$.

We denote the sample analogs of the above matrices by using the "hat" notation: $\hat{\mathbf{U}}_\mathbf{1}, \hat{\mathbf{U}}_\mathbf{0}, \hat{\mathbf{V}}_\mathbf{1}, \hat{\mathbf{V}}_\mathbf{0}, \hat{\mathbf{D}}_\mathbf{1}, \hat{\mathbf{D}}_\mathbf{0}$. These are generated by performing a singular value decomposition of $\mathbf{K_n}$.

Our test statistic estimates the sum of the squared elements of $\mathbf{D_0}$ (scaled by the total sample size, we will explain this later). We will now show how to diagonalize $\mathbf{K}$

to obtain $\mathbf{D_0}$.

$$\begin{aligned}
&\mathbf{U_0'KV_0} \\
&= \begin{bmatrix} \mathbf{U_0'U_1} & \mathbf{U_0'U_0} \end{bmatrix} \begin{bmatrix} \mathbf{D_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{D_0} \end{bmatrix} \begin{bmatrix} \mathbf{V_1'V_0} \\ \mathbf{V_0'V_0} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{U_0'U_1D_1} & \mathbf{U_0'U_0D_0} \end{bmatrix} \begin{bmatrix} \mathbf{V_1'V_0} \\ \mathbf{V_0'V_0} \end{bmatrix} \\
&= \mathbf{U_0'U_1D_1V_1'V_0 + U_0'U_0D_0V_0'V_0}
\end{aligned} \tag{4.4}$$

The vectors in $\mathbf{U_0}$ are orthogonal to those in $\mathbf{U_1}$. Identically, the vectors in $\mathbf{V_0}$ are orthogonal to those in $\mathbf{V_1}$. So the matrix products $\mathbf{U_0'U_1}$ and $\mathbf{V_1'V_0}$ are both zero. Also, the matrix product $\mathbf{U_0'U_0}$ contains elements that are dot products of vectors in $\mathbf{U_0}$. . The only non-zero dot products (again, because of the orthogonality) among these will be the dot products of a vector with itself. This will always be one (because of the orthonormality). The same situation occurs with $\mathbf{V_0'V_0}$.

$$\begin{aligned}
&\mathbf{U_0'KV_0} = \\
&\mathbf{0D_10 + I_kD_0I_h} = \\
&\mathbf{D_0}
\end{aligned} \tag{4.5}$$

So, our test statistic (4.2), estimates the sum of the squared elements of $\mathbf{U_0'KV_0}$. More precisely, the statistic estimates the following.

$$\begin{aligned}
&\left(\operatorname{vec}\left(\mathbf{U_0'KV_0}\right)\right)'\left(\operatorname{vec}\left(\mathbf{U_0'KV_0}\right)\right) = \\
&\left(\left(\mathbf{V_0'} \otimes \mathbf{U_0'}\right)\operatorname{vec}\left(\mathbf{K}\right)\right)'\left(\left(\mathbf{V_0'} \otimes \mathbf{U_0'}\right)\operatorname{vec}\left(\mathbf{K}\right)\right)
\end{aligned} \tag{4.6}$$

The singular value decomposition algorithm outputs a continuous function of the elements of its input matrix. Hence, if the input matrix $\mathbf{K_n}$ converges in probability, its singular vectors and values also converge in probability by the continuous mapping theorem. As explained in section 1.2.1 of Serfling (1980), our matrix $\mathbf{K_n}$ converges in

probability if and only if each of its components converges in probability.

By (3.82), for each component k_{nij} of $\mathbf{K_n}$, we have the following.

$$\sqrt{n}\left(k_{nij} - k_{ij}\right) \rightarrow^{d} N\left(0, \delta_{ij}\right) \tag{4.7}$$

Here the k_{ij} mean parameter is the ij component of $\mathbf{K}$ and the δ_{ij} variance parameter is the ij component of the variance matrix $\mathbf{\Delta}$ in (3.82). Because of this asymptotic distributional result, problem 1.P.20 in Serfling (1980) tells us that $k_{nij} \rightarrow^{p} k_{ij}$. Thus we conclude that $\mathbf{K_n} \rightarrow^{p} \mathbf{K}$. It immediately follows that $\mathbf{\hat{U}_1}, \mathbf{\hat{U}_0}, \mathbf{\hat{V}_1}, \mathbf{\hat{V}_0}, \mathbf{\hat{D}_1}, \mathbf{\hat{D}_0}$ each converge in probability via the continuous mapping theorem.

When $d = i$, Bura and Pfeiffer (2008) show that $\mathbf{\hat{U}_1}$ and $\mathbf{\hat{V}_1}$ converge in probability to $\mathbf{U_1}$ and $\mathbf{V_1}$. Hence $\mathbf{\hat{U}_1'K_n\hat{V}_1} = \mathbf{\hat{D}_1} \rightarrow^{p} \mathbf{D_1}$ by the continuous mapping theorem.

Under $d = i$, the matrices $\mathbf{U_0}$ and $\mathbf{V_0}$ do not necessarily seem to be uniquely defined. Returning to (4.4), if we examine the expressions without the pre and post multiplications we find the following.

$$\begin{aligned}
&\mathbf{K} \\
&= \begin{bmatrix} \mathbf{U_1} & \mathbf{U_0} \end{bmatrix} \begin{bmatrix} \mathbf{D_1} & \mathbf{0} \\ \mathbf{0} & \mathbf{D_0} \end{bmatrix} \begin{bmatrix} \mathbf{V_1'} \\ \mathbf{V_0'} \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{U_1D_1} & \mathbf{U_0D_0} \end{bmatrix} \begin{bmatrix} \mathbf{V_1'} \\ \mathbf{V_0'} \end{bmatrix} \\
&= \mathbf{U_1D_1V_1'} + \mathbf{U_0D_0V_0'} \\
&< d = i > \\
&= \mathbf{U_1D_1V_1'}
\end{aligned} \tag{4.8}$$

So when $d = i$, any $k - i$ vectors that are normalized and mutually orthogonal with each other and to the first i vectors in $\mathbf{U_1}$ may be in $\mathbf{U_0}$. Similarly, any $h - i$

vectors that are normalized and mutually orthogonal with each other and to the first i vectors in $\mathbf{V_1}$ may be in $\mathbf{V_0}$.

These restrictions define a class of vector sets that are suitable to be in $\mathbf{U_0}$, and a class of vector sets that are suitable to be in $\mathbf{V_0}$.

$$C_{\mathbf{U_0}}(\mathbf{U_1}) = \left\{ \begin{array}{l} A \ni \\ |A| = k - i; \\ \mathbf{x}, \mathbf{y} \in A \Rightarrow \\ \mathbf{x}'\mathbf{x} = 1, \mathbf{x}'\mathbf{y} = 0, \mathbf{x}'\mathbf{U_1} = \mathbf{0} \end{array} \right\} \tag{4.9}$$

$$C_{\mathbf{V_0}}(\mathbf{V_1}) = \left\{ \begin{array}{l} A \ni \\ |A| = h - i; \\ \mathbf{x}, \mathbf{y} \in A \Rightarrow \\ \mathbf{x}'\mathbf{x} = 1, \mathbf{x}'\mathbf{y} = 0, \mathbf{x}'\mathbf{V_1} = \mathbf{0} \end{array} \right\}$$

The vectors in $\hat{\mathbf{U}}_\mathbf{0}$ are in $C_{\mathbf{U_0}}\left(\hat{\mathbf{U}}_\mathbf{1}\right)$, and the criteria for membership in $C_{\mathbf{U_0}}\left(\hat{\mathbf{U}}_\mathbf{1}\right)$ are all based on continuous operations on the vectors in $\hat{\mathbf{U}}_\mathbf{0}$ and $\hat{\mathbf{U}}_\mathbf{1}$. It thus follows by the continuous mapping theorem (since $\hat{\mathbf{U}}_\mathbf{0}$ converges in probability to something and $\hat{\mathbf{U}}_\mathbf{1} \to^p \mathbf{U_1}$), that the vectors in $\hat{\mathbf{U}}_\mathbf{0}$ converge to an element of $C_{\mathbf{U_0}}(\mathbf{U_1})$. Similarly, the vectors in $\hat{\mathbf{V}}_\mathbf{0}$ converge to an element of $C_{\mathbf{V_0}}\left(\hat{\mathbf{V}}_\mathbf{1}\right)$. We can use the unique values that $\hat{\mathbf{U}}_\mathbf{0}$ and $\hat{\mathbf{V}}_0$ converge to as the theoretical singular vector parameters without any difficulties. For simplicity we say that under $d = i$ we have the following.

$$\begin{array}{l} \hat{\mathbf{U}}_\mathbf{0} \to^p \mathbf{U_0} \\ \hat{\mathbf{V}}_0 \to^p \mathbf{V_0} \end{array} \tag{4.10}$$

We can rewrite our test statistic in a similar manner to (4.6).

$$\begin{aligned}&\hat{\Lambda}_i = n \sum_{j=i+1}^{k} \hat{\lambda}_j = \\ &n \left(\left(\hat{\mathbf{V}}_{\mathbf{0}}' \otimes \hat{\mathbf{U}}_{\mathbf{0}}' \right) \mathrm{vec}\left(\mathbf{K_n} \right) \right)' \left(\left(\hat{\mathbf{V}}_{\mathbf{0}}' \otimes \hat{\mathbf{U}}_{\mathbf{0}}' \right) \mathrm{vec}\left(\mathbf{K_n} \right) \right)\end{aligned} \tag{4.11}$$

We do not immediately know the distribution of this reformulated expression, under $\mathrm{H}_0 : d = i$ or otherwise. We can find the distribution of a similar expression under $\mathrm{H}_0 : d = i$ and show that they are equivalent.

B. Mean Adjusted Test Statistic

This similar expression uses $\mathbf{K_n} - \mathbf{K}$ rather than just $\mathbf{K_n}$. Additionally, we will only need to use the parameter singular matrices, $\mathbf{U_0}$ and $\mathbf{V_0}$. rather than their estimates. It is quickly apparent that we can use the distribution of $\sqrt{n}\left(\mathbf{K_n} - \mathbf{K}\right)$ from (3.82).

By (3.82), under $\mathrm{H}_0 : d = i$ we have the following.

$$\begin{aligned}&\left(\mathbf{V}_{\mathbf{0}}' \otimes \mathbf{U}_{\mathbf{0}}'\right) \mathrm{vec}\sqrt{n}\left(\mathbf{K_n} - \mathbf{K}\right) \\ &\rightarrow^{d} \left(\mathbf{V}_{\mathbf{0}}' \otimes \mathbf{U}_{\mathbf{0}}'\right) N\left(\mathbf{0}, \boldsymbol{\Delta}\right) \\ &=^{d} N\left(\mathbf{0}, \left(\mathbf{V}_{\mathbf{0}}' \otimes \mathbf{U}_{\mathbf{0}}'\right) \boldsymbol{\Delta} \left(\mathbf{V}_{\mathbf{0}}' \otimes \mathbf{U}_{\mathbf{0}}'\right)'\right) \\ &=^{d} N\left(\mathbf{0}, \left(\mathbf{V}_{\mathbf{0}}' \otimes \mathbf{U}_{\mathbf{0}}'\right) \boldsymbol{\Delta} \left(\mathbf{V_0} \otimes \mathbf{U_0}\right)\right)\end{aligned} \tag{4.12}$$

So our test statistic is similar to the quadratic form obtained by diagonalizing the mean adjusted spanset estimate (using the parameters instead of their estimates for the diagonalizing matrices) and then squaring it. The expression in (4.11) is a quadratic form with each of the two input terms similar to (4.12). We now see how the sample size scaling factor becomes relevant. It is necessary to adjust the test statistic so that the variance is not infinite. This is done by scaling each term of the quadratic form. We want to obtain the distribution of a quadratic form of (4.11) under $\mathrm{H}_0 : d = i$. Noting that the dot product is a continuous function of its inputs,

by (4.11) and the continuous mapping theorem for distributional convergence, we have the following.

$$\begin{aligned}&\left((\mathbf{V_0'}\otimes\mathbf{U_0'})\,\mathrm{vec}\sqrt{n}\,(\mathbf{K_n}-\mathbf{K})\right)'\left((\mathbf{V_0'}\otimes\mathbf{U_0'})\,\mathrm{vec}\sqrt{n}\,(\mathbf{K_n}-\mathbf{K})\right)\\&\to^d \mathbf{y'y} \text{ where } \mathbf{y}\sim N\left(\mathbf{0},(\mathbf{V_0'}\otimes\mathbf{U_0'})\mathbf{\Delta}(\mathbf{V_0}\otimes\mathbf{U_0})\right)\end{aligned} \tag{4.13}$$

Now we can use a theorem adapted from page 113 of Eaton (1983) to determine this distribution. Another version occurs in Guttman (1982), but the earlier theorem needs a positive definite covariance matrix.

$$\begin{aligned}&\textbf{Quadratic Form Theorem}\\&\text{Let } \mathbf{y}\sim \mathrm{N_m}\left(\mathbf{0},\mathbf{\Sigma}\right),\mathbf{\Sigma} \text{ positive semidefinite}\\&\text{with eigenvalues } \lambda_1\geq\ldots\geq\lambda_m\\&\text{then } \mathbf{y'y} \text{ is distributed as } \textstyle\sum_{i=1}^m \lambda_i\chi_i\\&\chi_i\sim \text{i.i.d. } \chi_1^2\end{aligned} \tag{4.14}$$

So we conclude the following under $\mathrm{H_0}: d=i$.

$$\begin{aligned}&\left((\mathbf{V_0'}\otimes\mathbf{U_0'})\,\mathrm{vec}\sqrt{n}\,(\mathbf{K_n}-\mathbf{K})\right)'\left((\mathbf{V_0'}\otimes\mathbf{U_0'})\,\mathrm{vec}\sqrt{n}\,(\mathbf{K_n}-\mathbf{K})\right)\\&\to^d \textstyle\sum_{j=1}^{(k-i)(h-i)}\lambda_j\chi_j\\&\chi_j\sim \text{i.i.d. } \chi_1^2\\&\lambda_1\geq\ldots\geq\lambda_{(k-i)(h-i)} \text{ eigenvalues of } (\mathbf{V_0'}\otimes\mathbf{U_0'})\mathbf{\Delta}(\mathbf{V_0}\otimes\mathbf{U_0})\end{aligned} \tag{4.15}$$

By the continuous mapping theorem ($\otimes$ only multiplies and adds components from its argument matrices) and (4.10), under $\mathrm{H_0}: d=i$ we have the following.

$$\mathbf{\hat{V}_0'}\otimes\mathbf{\hat{U}_0'}\to^p\mathbf{V_0'}\otimes\mathbf{U_0'} \tag{4.16}$$

We can obtain a consistent estimator of $\mathbf{\Delta}$, $\mathbf{\hat{\Delta}}$ by using sample moment estimators (up to the fourth moment and including product moments) and the continuity mapping theorem. In the previous chapter we found that $\mathbf{\Delta}$ was quite complicated,

but its component pieces are all made of elementary algebraic manipulations of moments. Then we obtain the following, by the continuous mapping theorem and (4.16), under $H_0 : d = i$.

$$(\hat{\mathbf{V}}_\mathbf{0}' \otimes \hat{\mathbf{U}}_\mathbf{0}')\hat{\mathbf{\Delta}}(\hat{\mathbf{V}}_\mathbf{0} \otimes \hat{\mathbf{U}}_\mathbf{0}) \to^p (\mathbf{V}_\mathbf{0}' \otimes \mathbf{U}_\mathbf{0}')\mathbf{\Delta}(\mathbf{V}_\mathbf{0} \otimes \mathbf{U}_\mathbf{0}) \tag{4.17}$$

Via Lemma 2.1 from Tyler (1981) and (4.17), under $H_0 : d = i$ we obtain the following.

$$\begin{aligned} &i\text{th eigenvalue from } (\hat{\mathbf{V}}_\mathbf{0}' \otimes \hat{\mathbf{U}}_\mathbf{0}')\hat{\mathbf{\Delta}}(\hat{\mathbf{V}}_\mathbf{0} \otimes \hat{\mathbf{U}}_\mathbf{0}) \\ &\hat{\lambda}_i \to^p \lambda_i \\ &i\text{th eigenvalue from}(\mathbf{V}_\mathbf{0}' \otimes \mathbf{U}_\mathbf{0}')\mathbf{\Delta}(\mathbf{V}_\mathbf{0} \otimes \mathbf{U}_\mathbf{0}) \end{aligned} \tag{4.18}$$

So by Slutsky's Theorem, under $H_0 : d = i$ we have the following.

$$\begin{aligned} &\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_{(k-i)(h-i)} \text{ eigenvalues of } (\hat{\mathbf{V}}_\mathbf{0}' \otimes \hat{\mathbf{U}}_\mathbf{0}')\hat{\mathbf{\Delta}}(\hat{\mathbf{V}}_\mathbf{0} \otimes \hat{\mathbf{U}}_\mathbf{0}) \\ &\textstyle\sum_{j=1}^{(k-i)(h-i)} \hat{\lambda}_j \chi_j \to^d \sum_{m=1}^{(k-i)(h-i)} \lambda_m \chi_m \\ &\chi_j, \chi_m \sim \text{i.i.d. } \chi_1^2 \\ &\lambda_1 \geq \ldots \geq \lambda_{(k-i)(h-i)} \text{ eigenvalues of } (\mathbf{V}_\mathbf{0}' \otimes \mathbf{U}_\mathbf{0}')\mathbf{\Delta}(\mathbf{V}_\mathbf{0} \otimes \mathbf{U}_\mathbf{0}) \end{aligned} \tag{4.19}$$

So we conclude the following under $H_0 : d = i$.

$$\begin{aligned} &((\mathbf{V}_\mathbf{0}' \otimes \mathbf{U}_\mathbf{0}')\operatorname{vec}\sqrt{n}\,(\mathbf{K_n}-\mathbf{K}))' ((\mathbf{V}_\mathbf{0}' \otimes \mathbf{U}_\mathbf{0}')\operatorname{vec}\sqrt{n}\,(\mathbf{K_n}-\mathbf{K})) \\ &\approx^d \\ &\textstyle\sum_{j=1}^{(k-i)(h-i)} \hat{\lambda}_j \chi_j \\ &\chi_j \sim \text{i.i.d. } \chi_1^2 \\ &\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_{(k-i)(h-i)} \text{ eigenvalues of } (\hat{\mathbf{V}}_\mathbf{0}' \otimes \hat{\mathbf{U}}_\mathbf{0}')\hat{\mathbf{\Delta}}(\hat{\mathbf{V}}_\mathbf{0} \otimes \hat{\mathbf{U}}_\mathbf{0}) \end{aligned} \tag{4.20}$$

So we can use our estimated covariance matrix of the spanning set together with the estimated singular vectors from the spanning set to estimate the null distribution of (4.13). It is time to show how (4.13) and our test statistic are equivalent under

$H_0 : d = i.$

C. Equivalence Mean Adjusted Test Statistic and $\hat{\Lambda}_i$

As mentioned before, the singular value decomposition is a continuous function of the input matrix elements. For $q \leq p$, let $\psi : \mathbb{R}^{p\times q} \mapsto \mathbb{R}^q$ be the singular value function, receiving real valued matrices as input and outputting a vector of the input's singular values. In our situation, the matrices $\mathbf{K_n}$ and $\mathbf{K}$ have more columns than rows, so we could pass their transposes to ψ.

In Eaton and Tyler (1994), the asymptotic distribution of $\sqrt{n}\left(\psi\left(\mathbf{X_n}\right) - \psi\left(\mathbf{B}\right)\right)$ was studied, under the assumption that $\sqrt{n}\left(\mathbf{X_n} - \mathbf{B}\right)$ converged to a valid distribution. Let's apply this notation to our setting under $H_0 : d = i$.

$$\sqrt{n}\left(\psi\left(\mathbf{K_n}\right) - \psi\left(\mathbf{K}\right)\right) =$$
$$\sqrt{n}\left(\begin{bmatrix} \operatorname{diag}(\hat{\mathbf{U}}_\mathbf{1}'\mathbf{K_n}\hat{\mathbf{V}}_\mathbf{1}) & = \operatorname{diag}(\hat{\mathbf{D}}_\mathbf{1}) = & \begin{matrix}\hat{\sigma}_1 \\ \vdots \\ \hat{\sigma}_i\end{matrix} \\ \operatorname{diag}(\hat{\mathbf{U}}_\mathbf{0}'\mathbf{K_n}\hat{\mathbf{V}}_\mathbf{0}) & = \operatorname{diag}(\hat{\mathbf{D}}_\mathbf{0}) & \end{bmatrix} - \begin{bmatrix} \sigma_1 \\ \vdots \\ \sigma_i \\ \mathbf{0_{min(k-i,h-i)}} \end{bmatrix}\right) \tag{4.21}$$

They concluded, that under $H_0 : d = i$, the distribution of the components of (4.21) indexed above i had the same asymptotic distribution as $\sqrt{n}\psi\left(\mathbf{U_0'K_nV_0}\right)$. Recalling (4.8) and the orthogonality of the 0 partitions with their 1 partition counterparts, this asymptotic distribution is equivalent to $\sqrt{n}\psi\left(\mathbf{U_0'}\left(\mathbf{K_n} - \mathbf{K}\right)\mathbf{V_0}\right)$

So we conclude the following under $H_0 : d = i$.

$$\begin{aligned} &\sqrt{n}\left(\operatorname{diag}(\hat{\mathbf{U}}_\mathbf{0}'\mathbf{K_n}\hat{\mathbf{V}}_\mathbf{0})\right) \approx^d \\ &\sqrt{n}\psi(\mathbf{U_0'}\left(\mathbf{K_n} - \mathbf{K}\right)\mathbf{V_0}) \end{aligned} \tag{4.22}$$

Now we obtain the following by noting that $\hat{\mathbf{U}}_\mathbf{0}'\mathbf{K_n}\hat{\mathbf{V}}_\mathbf{0}$ is diagonal and that we

can pull the $\left(\hat{\mathbf{V}}_\mathbf{0}' \otimes \hat{\mathbf{U}}_\mathbf{0}'\right)$ term back into the vec.

$$\begin{aligned}
&\sqrt{n}\left(\text{diag}(\hat{\mathbf{U}}_\mathbf{0}'\mathbf{K_n}\hat{\mathbf{V}}_\mathbf{0})\right)'\sqrt{n}\left(\text{diag}(\hat{\mathbf{U}}_\mathbf{0}'\mathbf{K_n}\hat{\mathbf{V}}_\mathbf{0})\right) = \\
&n\left(\text{vec}\left(\hat{\mathbf{U}}_\mathbf{0}'\mathbf{K_n}\hat{\mathbf{V}}_\mathbf{0}\right)\right)'\left(\text{vec}\left(\hat{\mathbf{U}}_\mathbf{0}'\mathbf{K_n}\hat{\mathbf{V}}_\mathbf{0}\right)\right) = \\
&n\left(\left(\hat{\mathbf{V}}_\mathbf{0}' \otimes \hat{\mathbf{U}}_\mathbf{0}'\right)\text{vec}\left(\mathbf{K_n}\right)\right)'\left(\left(\hat{\mathbf{V}}_\mathbf{0}' \otimes \hat{\mathbf{U}}_\mathbf{0}'\right)\text{vec}\left(\mathbf{K_n}\right)\right)
\end{aligned} \tag{4.23}$$

This is our test statistic from written in the form of (4.11). By the continuous mapping theorem and the above, under $\text{H}_0 : d = i$ we obtain the following result.

$$\begin{aligned}
&n\left(\left(\hat{\mathbf{V}}_\mathbf{0}' \otimes \hat{\mathbf{U}}_\mathbf{0}'\right)\text{vec}\left(\mathbf{K_n}\right)\right)'\left(\left(\hat{\mathbf{V}}_\mathbf{0}' \otimes \hat{\mathbf{U}}_\mathbf{0}'\right)\text{vec}\left(\mathbf{K_n}\right)\right) \approx^d \\
&\sqrt{n}\psi(\mathbf{U}_\mathbf{0}'\left(\mathbf{K_n} - \mathbf{K}\right)\mathbf{V_0})'\sqrt{n}\psi(\mathbf{U}_\mathbf{0}'\left(\mathbf{K_n} - \mathbf{K}\right)\mathbf{V_0})
\end{aligned} \tag{4.24}$$

The latter half of (4.24) is the sum of the squared singular values of the matrix $\sqrt{n}\mathbf{U}_\mathbf{0}'\left(\mathbf{K_n} - \mathbf{K}\right)\mathbf{V_0}$. These are the eigenvalues of the matrix $\left(\sqrt{n}\mathbf{U}_\mathbf{0}'\left(\mathbf{K_n} - \mathbf{K}\right)\mathbf{V_0}\right)\left(\sqrt{n}\mathbf{U}_\mathbf{0}'\left(\mathbf{K_n} - \mathbf{K}\right)\mathbf{V_0}\right)'$. The sum of these eigenvalues is equal to the trace of this matrix.

There is a useful relationship between the vec and trace operators. For matrices $\mathbf{A}$ and $\mathbf{B}$ such that $\mathbf{BA}$ (and therefore $\mathbf{AB}$) is square, $\text{trace}\left(\mathbf{BA}\right) = \text{vec}\left(\mathbf{A}'\right)'\text{vec}\left(\mathbf{B}\right)$. This result and other useful matrix algebra identities are found in the appendix of Lütkepohl (2007).

We take advantage of this relationship between the vec and trace operators to obtain the following.

$$\begin{aligned}
&\sqrt{n}\psi(\mathbf{U}_\mathbf{0}'\left(\mathbf{K_n} - \mathbf{K}\right)\mathbf{V_0})'\sqrt{n}\psi(\mathbf{U}_\mathbf{0}'\left(\mathbf{K_n} - \mathbf{K}\right)\mathbf{V_0}) = \\
&\text{trace}\left(\sqrt{n}\mathbf{U}_\mathbf{0}'\left(\mathbf{K_n} - \mathbf{K}\right)\mathbf{V_0}\left(\sqrt{n}\mathbf{U}_\mathbf{0}'\left(\mathbf{K_n} - \mathbf{K}\right)\mathbf{V_0}\right)'\right) = \\
&n\left(\text{vec}\left(\mathbf{U}_\mathbf{0}'\left(\mathbf{K_n} - \mathbf{K}\right)\mathbf{V_0}\right)\right)'\left(\text{vec}\left(\mathbf{U}_\mathbf{0}'\left(\mathbf{K_n} - \mathbf{K}\right)\mathbf{V_0}\right)\right) = \\
&n\left(\left(\mathbf{V}_\mathbf{0}' \otimes \mathbf{U}_\mathbf{0}'\right)\text{vec}\left(\mathbf{K_n} - \mathbf{K}\right)\right)'\left(\left(\mathbf{V}_\mathbf{0}' \otimes \mathbf{U}_\mathbf{0}'\right)\text{vec}\left(\mathbf{K_n} - \mathbf{K}\right)\right)
\end{aligned} \tag{4.25}$$

Therefore we conclude the following by (4.24) and (4.11).

$$\begin{aligned} & n\left(\left(\mathbf{V_0'}\otimes\mathbf{U_0'}\right)\mathrm{vec}\left(\mathbf{K_n}-\mathbf{K}\right)\right)'\left(\left(\mathbf{V_0'}\otimes\mathbf{U_0'}\right)\mathrm{vec}\left(\mathbf{K_n}-\mathbf{K}\right)\right)\approx^{d} \\ & n\left(\left(\hat{\mathbf{V}}_{\mathbf{0}}'\otimes\hat{\mathbf{U}}_{\mathbf{0}}'\right)\mathrm{vec}\left(\mathbf{K_n}\right)\right)'\left(\left(\hat{\mathbf{V}}_{\mathbf{0}}'\otimes\hat{\mathbf{U}}_{\mathbf{0}}'\right)\mathrm{vec}\left(\mathbf{K_n}\right)\right)= \\ & n\sum_{j=i+1}^{k}\hat{\lambda}_j=\hat{\Lambda}_i \end{aligned} \tag{4.26}$$

D. 3 Reference Distributions of $\hat{\Lambda}_i$

So by (4.15), under $\mathrm{H}_0: d=i$ our test statistic from (4.2) has the following asymptotic distribution.

Theory

$$\begin{aligned} & \hat{\Lambda}_i \\ & \rightarrow^{d}\textstyle\sum_{j=1}^{(k-i)(h-i)}\lambda_j\chi_j \\ & \chi_j\sim \text{i.i.d. } \chi_1^2 \\ & \lambda_1\geq\ldots\geq\lambda_{(k-i)(h-i)} \text{ eigenvalues of } (\mathbf{V_0'}\otimes\mathbf{U_0'})\mathbf{\Delta}(\mathbf{V_0}\otimes\mathbf{U_0}) \end{aligned} \tag{4.27}$$

This result is theoretically useful, but it does not have practical merit. To perform the dimensionality test in practice, one would have to already know the diagonalized spanset variance matrix parameter or its eigenvalues.

Earlier we showed how the estimates of these parameters converged in probability. We demonstrated how the distribution of the quadratic form of diagonalized and mean adjusted statistic (4.13) could be approximated using these estimates in (4.20). Given

(4.26), under $H_0 : d = i$ we obtain the following distribution.

$$\begin{aligned}
&\textbf{Empirical} \\
&\hat{\Lambda}_i \approx^d \\
&\textstyle\sum_{j=1}^{(k-i)(h-i)} \hat{\lambda}_j \chi_j \\
&\chi_j \sim \text{i.i.d. } \chi_1^2 \\
&\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_{(k-i)(h-i)} \text{ eigenvalues of } (\hat{\mathbf{V}}_\mathbf{0}' \otimes \hat{\mathbf{U}}_\mathbf{0}')\hat{\mathbf{\Delta}}(\hat{\mathbf{V}}_\mathbf{0} \otimes \hat{\mathbf{U}}_\mathbf{0})
\end{aligned} \tag{4.28}$$

So we can perform the tests in (4.1) using only the sample data. We will find that using (4.28) to directly perform the dimensionality test of $H_0 : d = i$ will require a simulation.

Unlike other more simple distributions, such as a chi-squared or normal distribution, the quantiles of the distribution in (4.28) cannot be automatically calculated. The cumulative distribution function and the density of (4.28) are not explicitly known. A full derivation of their form for a given spanning set could be performed, but it will be very complicated.

We can perform a simulation to estimate the cumulative distribution function of the right hand side of (4.28) and its quantiles. We would compare our test statistic to these quantiles and then accept $H_0 : d = i$ unless the statistic was close to or exceeding the estimated high quantiles. Equivalently we could count the number of times that our simulated value of (4.28) exceeded the value of the test statistic. The ratio of this count with the total number of simulated values of (4.28) would yield a p-value for the test $H_0 : d = i$.

To perform the simulation, we first calculate $(\hat{\mathbf{V}}_\mathbf{0}' \otimes \hat{\mathbf{U}}_\mathbf{0}')\hat{\mathbf{\Delta}}(\hat{\mathbf{V}}_\mathbf{0} \otimes \hat{\mathbf{U}}_\mathbf{0})$ and its eigenvalues. Every realization of the random variable in (4.28) is a dot product of a vector containing these values with a vector of independent χ_1^2. Individual χ_1^2 realizations are simple to produce. So after the calculation of the parameters, realizations

of the distribution of (4.28) are easy to make.

In the simulation, a large number of independent realizations are generated. The quantiles are then estimated, treating the entire set of realizations as if it were the entire support of the distribution in (4.28).

This method of testing by running a simulation is effective, but it is not very fast. A user would prefer to be able to automatically determine the quantiles (as they can when performing a T-test for example) of the test statistic's null distribution and then automatically obtain an accept/reject decision. This concern is well founded. The other inverse regression techniques, SIR and SAVE have encountered this issue as well. One solution is to approximate the linear combination of χ^2_1 random variables with a single "corrected" chi-squared random variable. This was explained and studied in Bentler and Xie (2000). The new random variable is scaled and its degree of freedom are adjusted to reflect the affect of the eigenvalue coefficients. The R language dr package (Weisberg (2009)) implementation of SIR and SAVE allows the use of this method.

We will apply this method to approximating the distribution 4.28 in our test of $\mathrm{H}_0 : d = i$. First we formally state the result we will use from Bentler and Xie (2000).

Suppose a statistic $T \rightarrow^d \sum_{i=1}^m \alpha_i \chi^2_1$ where the χ^2_1 variables are independent and α_i coefficients are the eigenvalues of a symmetric non-negative definite matrix $\mathbf{A}$. We can estimate $\mathbf{A}$ with a consistent (convergent in probability) estimate matrix $\mathbf{A_n}$. Then we have the following result.

$$\begin{aligned} &\text{For } d \text{ closest integer to } \tfrac{\mathrm{trace}(\mathbf{A_n})^2}{\mathrm{trace}(\mathbf{A_n^2})} \\ &\overline{\overline{T}} = \tfrac{d}{\mathrm{trace}(\mathbf{A_n})} T \approx {}^d\chi^2_d \end{aligned} \tag{4.29}$$

Using (4.27), this applies to our situation. We can obtain a consistent estimate of our $\mathbf{A}$ matrix $(\mathbf{V_0'} \otimes \mathbf{U_0'})\boldsymbol{\Delta}(\mathbf{V_0} \otimes \mathbf{U_0})$ in $(\hat{\mathbf{V}}_\mathbf{0}' \otimes \hat{\mathbf{U}}_\mathbf{0}')\hat{\boldsymbol{\Delta}}(\hat{\mathbf{V}}_\mathbf{0} \otimes \hat{\mathbf{U}}_\mathbf{0})$ via (4.17). Note that

$\mathbf{A}$ must be non-negative definite. Our $(\mathbf{V_0'} \otimes \mathbf{U_0'})\mathbf{\Delta}(\mathbf{V_0} \otimes \mathbf{U_0})$ is a variance matrix (as demonstrated in (4.13)) so it is certainly non-negative definite.

Approximate Empirical

$$\text{For } d \text{ closest integer to } \frac{\text{trace}\left((\hat{\mathbf{V}}_\mathbf{0}' \otimes \hat{\mathbf{U}}_\mathbf{0}')\hat{\mathbf{\Delta}}(\hat{\mathbf{V}}_\mathbf{0} \otimes \hat{\mathbf{U}}_\mathbf{0})\right)^2}{\text{trace}\left(\left((\hat{\mathbf{V}}_\mathbf{0}' \otimes \hat{\mathbf{U}}_\mathbf{0}')\hat{\mathbf{\Delta}}(\hat{\mathbf{V}}_\mathbf{0} \otimes \hat{\mathbf{U}}_\mathbf{0})\right)^2\right)} \quad \frac{d}{\text{trace}\left((\hat{\mathbf{V}}_\mathbf{0}' \otimes \hat{\mathbf{U}}_\mathbf{0}')\hat{\mathbf{\Delta}}(\hat{\mathbf{V}}_\mathbf{0} \otimes \hat{\mathbf{U}}_\mathbf{0})\right)}\hat{\Lambda}_i \approx {}^d\chi^2_d \tag{4.30}$$

So a user of the SMVCIR dimensionality test may use the empirical (4.28) or approximate empirical (4.30). We will compare both approaches. To corroborate our results and show how effective our testing method can be, we will perform simulations and go through several examples. The approximate empirical (4.30), empirical (4.28), and theory (4.27) distribution results will be studied.

CHAPTER V

CALCULATION/TEST DETAILS AND A COMPLETE EXAMPLE

We have provided a lot of theory in the last two chapters. It will be better explained when we apply it in an example. A more thorough explanation of how we would calculate some of the quantities in our theory is also necessary. Additionally, before simulations are performed in the next chapter, we need to outline some of our simulation evaluation methodology.

We would like to be able to see how the test statistic behaves over multiple samples. This can be done in several ways using the approximate empirical and empirical distribution results. For the empirical, we may form a rejection rule based on the simulated quantiles and report the type 1 error rate of the test overall the samples. We would conclude that the test performed well in the given situation if the probability for the quantile we used for the rejection rule matched the value of the error rate. We could perform a similar exercise to examine the approximate empirical reference distribution, but here the quantiles would be automatic and require no simulation.

We will perform both these examinations in a variety of contexts. Examination of the power of the tests under the alternative hypotheses will also be conducted. Particularly, we will compare the tests in both power and size across a variety of different sample sizes and population settings. The empirical will be compared with the empirical across different sample sizes and population settings The approximate empirical will be compared with the approximated empirical across different sample sizes and population settings. And both test forms will be compared with each other across different sample sizes and population settings.

As explained in the last chapter, we can estimate the $\mathbf{\Delta}$ matrix using sample

moments and proportions. The $\mathbf{K}$ spanning set matrix can be well estimated by $\mathbf{K_n}$ which is easily calculated using the sample means, variances, covariances, and proportions, so its singular value decomposition matrices are easily estimated as well. So the estimated covariance matrix of the diagonalize spanning set estimate used in the approximate empirical (4.30) and empirical (4.28) distributions is easily calculated using the data. So we can perform the tests in our simulations without extra considerations about the SMVCIR parameters.

But we should realize that the power and size of our tests that we observe in simulations, are themselves statistics. Their probabilistic behavior should be considered, particularly the situations where their behavior drastically changes. The power of the tests compared may be drastically different depending on the actual size of the test. When we supposedly fix a size for our tests α, the actual size $\alpha^\star$ may differ. This means that the power may not be simply and directly controlled by the nominal size of the test α. In our situation we may likely encounter this in small sample sizes, since our distributional results are only valid in large samples. These issues were explored in Lloyd (2005). We incorporate Lloyd's "method 1" into our simulation studies. This allows us to correctly compare the power of tests, even if their actual sizes $\alpha^\star$ differ. The details will be found in the next section.

It is also of interest to examine the theory (4.27) null distribution of the dimensionality test statistic. In this case we can graphically examine how the test statistic behaves over multiple samples, by comparing its estimated distribution with that of the theory distribution. A scatter plot of the test statistic and theory quantiles (a Q-Q plot) may be used to compare the two. We will also compare the test statistic with the theory distribution by overlaying the kernel density estimates of the two.

To make these graphical comparisons, we have to create realizations from the theory distribution. So we will have to explicitly calculate the diagonalized variance

matrix $(\mathbf{V_0'} \otimes \mathbf{U_0'})\mathbf{\Delta}(\mathbf{V_0} \otimes \mathbf{U_0})$. This explicitly depends on the spanning set estimate variance matrix $\mathbf{\Delta}$ and implicitly depends on the spanning set $\mathbf{K}$ through the singular value decomposition. We would like to postulate group means and variances (both on the unstandardized population) and group proportions. Then we want to be able to derive all the parameters of the asymptotic mean and variance of our estimators with minimal extra assumptions.

In this chapter we first provide the details of Lloyd's "method 1" as it applies to our work here. In the next section we provide calculation details for the theoretical spanning set $\mathbf{K}$ and covariance of our dimensionality test statistic $\mathbf{\Delta}$. Finally in the last section we give two examples demonstrating how we perform SMVCIR and use the dimensionality test.

A. Test Power Comparison

The power of a test is formally the probability of rejecting the null hypothesis under a particular parameter value. For a size α test, the power of the test under each parameter value in H_0 should be less than or equal to α. Equivalently, α is defined as the supremum of the powers under H_0. The probability of correct decision to reject H_0 under a particular parameter value in the alternative hypothesis H_1 (here for a particular alternative value to d) is β. In Lloyd's notation, β is referred to as the "power" of the test.

In this section we draw a distinction between the nominal size of the test (the value used to formulate the rejection rule) α^{nom} and the true size of the test α. We also define separate powers β_i for each $i > d$.

We examine each test individually. In the last chapter we discuss how we might study the entire sequence of tests as a future research direction. In multiple hypothesis

test situations, the nominal size we use for each individual test α^{nom} may be quite different than the actual size of the combined tests together. The most familiar case of this happens with multiple comparison of the means in an ANOVA. Our situation, using the iterated test sequence in (4.1) is even more complicated, as inference only leads one to examine tests until the null is accepted. So each test is nested within the last. This form of testing is called sequential hypothesis testing or sequential analysis.

Lloyd's method of test power comparison (Lloyd (2005)) depends on having an estimator of the test size α, $\hat{\alpha}$ and an estimator of the power under the alternative $\hat{\beta}$. In SMVCIR, the dimensionality parameter d can conceivably take values $d = 0, 1, \ldots, k$. These different values for d are caused by different values for the group mean and variance/covariance parameters. So for the test statistic $\hat{\Lambda}_i$ with nominal test size α^{nom} for $\text{H}_0 : d = i$ the true size, α and power, β_i parameters are dictated by the continuous and multidimensional group mean and variance parameters.

We will evaluate $\hat{\Lambda}_i$ in a variety of situations (a variety of different group mean and variance parameters) for a variety of α^{nom} values. In each situation, we can easily estimate α for the test statistic $\hat{\Lambda}_d$ under the empirical or approximate empirical reference distributions.

For estimation of β_i we will formulate similar situations (same nominal power, same dimensionality, same group proportions) to the situation that gave us our α estimate. In each situation, the true dimensionality of the SMVCIR space will exceed $d = d_0$, the dimension for the test statistic we used for estimation of α. For each of the possible alternatives $i > d_0$, we will estimate $\hat{\beta}_i$ as the proportion of correct decisions to reject H_0 when the true $d = i$ using $\hat{\Lambda}_{d_0}$. Our estimation of α, and β_i will be by no means exhaustive, since there are so many possibilities for the population parameters. But they will be varied enough that we have some reasonable breadth in our contexts.

Once we have estimates $\hat{\alpha}$ and $\hat{\beta}_i$, we can compare any two tests of the same null hypothesis $d = d_0$ and alternative hypothesis parameter values $d = i > d_0$. We use Lloyd's method 1. We will also only compare tests that use identical α^{nom} values. Obviously comparison of a .1 level test and a .05 level test should lead to some differences. Also, the compared tests will obviously be sampled under identical circumstances as well.

Lloyd's method 1 estimates the relationship of α and β (where the subscript on β is understood) as being well approximated (when the number of sampled tests is large) by $\beta = \Phi\left(\delta + \Phi^{-1}\left(\alpha\right)\right)$. The Φ symbol represents the standard normal cumulative distribution function. The δ parameter represents the quality of the test and can be estimated by the following.

$$\hat{\delta} = \Phi^{-1}(\hat{\beta}) - \Phi^{-1}(\hat{\alpha}) \tag{5.1}$$

We can compare two tests by examining their $\hat{\delta}$ values. Of course, as we mentioned earlier, $\hat{\alpha}$ and $\hat{\beta}$ are statistics. So $\hat{\delta}$ is a statistic, and we can perform inference on the equivalence of the tests 1 and 2 using the statistical properties of $\hat{\delta}_1$ and $\hat{\delta}_2$.

We will use 6 criteria to compare tests. We have already discussed the δ criterion. Fixing the null and alternative hypotheses and the nominal size of the test α^{nom} for tests 1 and 2, it is also instructive to compare the estimates $\hat{\alpha}_1$ and $\hat{\alpha}_2$ and $\hat{\beta}_1$ and $\hat{\beta}_2$. So we will use α and β as criteria for test comparison. Their difference $\beta - \alpha$ will also be a useful criterion.

A value of β for the nominal alpha value α^{nom} is also a useful criteria. Following Lloyd's notation, we denote this value as the following.

$$R\left(\alpha^{\text{nom}}\right) = \Phi\left(\delta + \Phi^{-1}\left(\alpha^{\text{nom}}\right)\right) \tag{5.2}$$

The final criteria is the "proportion of correct classifications" parameter.

$$W(0,1) = \Phi\left(\frac{\delta}{\sqrt{2}}\right) \tag{5.3}$$

These last two criteria are estimated as follows.

$$\hat{R}\left(\alpha^{\text{nom}}\right) = \Phi\left(\hat{\delta} + \Phi^{-1}\left(\alpha^{\text{nom}}\right)\right) \tag{5.4}$$

$$\hat{W}(0,1) = \Phi\left(\frac{\hat{\delta}}{\sqrt{2}}\right) \tag{5.5}$$

So we compare tests of the same α^{nom} and null dimension d_0 against the same alternative dimension d_1 using the six criteria: true size α, power β, $\alpha - \beta$, δ, $R\left(\alpha^{\text{nom}}\right)$, and coefficient of correct classification $W(0,1)$. Each of the criteria have estimates that have been previously noted. Lloyd calculated the standard errors of these estimates as well. He refrained from developing their distributions in method 1.

We will compare tests using only the estimates of the criteria and their standard errors. Lloyd's other methods of test comparison provide full statistical inference on the criteria, but they are complex and involve time consuming techniques like the bootstrap. Performing each test is already a somewhat time consuming technique due to the complicated large matrix calculations. So, we will compare tests using accurate point estimates of comparison criteria, while having an estimate of how precise the point estimate is.

These are the estimated standard errors of the six criteria estimates. By "standard error", we mean asymptotic standard deviation of the estimate. The value B_0 is the number of tests sampled under the the null dimension d_0. The value B_1 is the number of tests sampled under the alternative dimension d_1. Lloyd explicitly calculated each, and we produce their estimates with the normal "plug-in" method.

Note that the the null and alternative tests are independently conducted, so $\hat{\alpha}$ and $\hat{\beta}$ are independent. Also we use ϕ to represent the standard normal density.

$$\hat{\mathrm{SD}}\left(\hat{\alpha}\right) = \sqrt{\frac{\hat{\alpha}\sqrt{1-\hat{\alpha}}}{B_0}} \tag{5.6}$$

$$\hat{\mathrm{SD}}\left(\hat{\beta}\right) = \sqrt{\frac{\hat{\beta}\sqrt{1-\hat{\beta}}}{B_1}} \tag{5.7}$$

$$\hat{\mathrm{SD}}\left(\hat{\beta}-\hat{\alpha}\right) = \sqrt{\hat{\mathrm{SD}}(\hat{\beta})^2 + \hat{\mathrm{SD}}\left(\hat{\alpha}\right)^2} \tag{5.8}$$

$$\hat{\mathrm{SD}}\left(\hat{\delta}\right) = \sqrt{\frac{\hat{\mathrm{SD}}(\hat{\beta})^2}{\phi^2(\Phi^{-1}(\hat{\beta}))} + \frac{\hat{\mathrm{SD}}\left(\hat{\alpha}\right)^2}{\phi^2\left(\Phi^{-1}\left(\hat{\alpha}\right)\right)}} \tag{5.9}$$

$$\hat{\mathrm{SD}}\left(\hat{R}\left(\alpha^{\mathrm{nom}}\right)\right) = \hat{\mathrm{SD}}\left(\hat{\delta}\right)\phi\left(\hat{\delta}+\Phi^{-1}\left(\alpha^{\mathrm{nom}}\right)\right) \tag{5.10}$$

$$\hat{\mathrm{SD}}\left(\hat{W}\left(0,1\right)\right) = \hat{\mathrm{SD}}\left(\hat{\delta}\right)\left(\frac{1}{\sqrt{2}}\right)\phi\left(\frac{\hat{\delta}}{\sqrt{2}}\right) \tag{5.11}$$

B. Calculation of $\mathbf{\Delta}$ and $\mathbf{K}$

In Chapter III we found that the asymptotic distribution of the spanning set estimate depended on the parameter matrices $\mathbf{K}$ and $\mathbf{\Delta}$. In the last chapter we found that the theory distribution depended on the the same matrices, $\mathbf{\Delta}$ directly and $\mathbf{K}$ through the singular value decomposition. These parameter matrices are functions of the original parameters of the asymptotic distribution of $\bar{\mathbf{u}}$ in (3.22).

In Chapter III, the mean of $\bar{\mathbf{u}}$ depended on the pseudo-first moment parameters $\mathbf{m_j}$, second moment parameters $\mathbf{V_j}$, group proportions p_j, marginal mean $\boldsymbol{\mu}$, and

marginal variance $\mathbf{\Sigma}$. In addition to these parameters the asymptotic variance of $\bar{\mathbf{u}}$ depends on the parameters $\mathbf{M_j} = \mathrm{E}[1_{ji}\mathbf{x_i}(\mathbf{x_i'} \otimes \mathbf{x_i'})]$ and $\mathbf{N_j} = \mathrm{E}\left[1_{ji}(\mathbf{x_i}\mathbf{x_i'}) \otimes (\mathbf{x_i}\mathbf{x_i'})\right]$, and their non-group counterparts $\mathbf{M} = \mathrm{E}[\mathbf{x_i}(\mathbf{x_i'} \otimes \mathbf{x_i'})]$ and $\mathbf{N} = \mathrm{E}\left[(\mathbf{x_i}\mathbf{x_i'}) \otimes (\mathbf{x_i}\mathbf{x_i'})\right]$.

Under our initial assumptions, the $\mathbf{m_j}$, $\mathbf{V_j}$, p_j, $\boldsymbol{\mu}$, and $\mathbf{\Sigma}$ parameters are easily obtained. The $\mathbf{M_j}$, $\mathbf{N_j}$, $\mathbf{M}$, and $\mathbf{N}$ parameters involve third and fourth moment parameters. We need additional assumptions to derive them.

The multivariate normal distribution is completely determined by its first and second moments. It also has some very useful properties that allow straightforward calculation of its higher order moments. Namely, all subvectors are multivariate normal and conditioning one subvector on another subvector leads to a multivariate normal. So we can calculate $\mathbf{M_j}$, $\mathbf{N_j}$, $\mathbf{M}$, and $\mathbf{N}$ with our initial mean and variance assumptions if we assume the underlying group populations are multivariate normal.

Using only multivariate normal populations is limiting. We can add more flexibility by considering mixtures of multivariate normals. This allows multi modality and a variety of other attributes (including skewness) in a group's distribution. By doing this we have two levels of mixing. At the highest level we have the individual groups that we will discriminate using SMVCIR. And then at intra-group level we may have multiple multivariate normal populations that we observe. Remember that SMVCIR only pays attention to the marginal mean and variance of the group (over all the multiple populations within the group). Also we retain our restriction that the means must be defined and the variance must be non-singular for each subpopulation within a group.

The calculation of the necessary parameters in this two level mixture case is straightforward. To calculate the group proportions, means, variances, and second-fourth order moments that determine the asymptotic distribution of $\bar{\mathbf{u}}$ and therefore determine $\mathbf{\Delta}$ and $\mathbf{K}$, we merely need to use two levels of conditioning in our calcu-

lations (instead of only one). So we can use the non-mixture multivariate normal formulae for calculation of the moments and then take care to properly condition on the group and properly marginalize over the sub-categories within that group.

So in evaluation of the theory (4.27) null distribution of the dimensionality test statistic we will assume that the underlying group populations are k-variate multivariate normal or mixture multivariate normal. As before there are g groups.

Suppose the group i comes from a mixture of g_i multivariate normals (we have the non-mixture case when $g_i = 1$). We index the individual population parameters within the group with a superscript to differentiate them from the group mean and its components. So the individual population mean of population $1 \leq j \leq g_i$ within group i is denoted $\boldsymbol{\mu}_{\mathbf{i}}^{\mathbf{j}}$. Similarly the individual population variance of population $1 \leq j \leq g_i$ within group i is denoted $\boldsymbol{\Sigma}_{\mathbf{i}}^{\mathbf{j}}$. The group i population proportion of population $1 \leq j \leq g_i$ within group i is denoted p_i^j.

Please recall our derivation of the marginal mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$ in terms of the group means, variances, and proportions in Chapter I. We can apply the same steps to calculating the marginal means and variances for group i over the $1, \ldots, g_i$ subpopulations. The group i population proportion p_i is still axiomatically defined and does not depend on the subpopulations.

So group i has the following mean and variance.

$$\boldsymbol{\mu}_{\mathbf{i}} = \sum_{j=1}^{g_i} p_i^j \boldsymbol{\mu}_{\mathbf{i}}^{\mathbf{j}} \tag{5.12}$$

$$\boldsymbol{\Sigma} = \sum_{j=1}^{g_i} \boldsymbol{\Sigma}_{\mathbf{i}}^{\mathbf{j}} p_i^j + \sum_{j=1}^{g_i} \boldsymbol{\mu}_{\mathbf{i}}^{\mathbf{j}} \boldsymbol{\mu}_{\mathbf{i}}^{\mathbf{j}'} p_i^j - \boldsymbol{\mu}_{\mathbf{i}} \boldsymbol{\mu}_{\mathbf{i}}' \tag{5.13}$$

Recall from (1.1) that the marginal mean is defined as the following.

$$\boldsymbol{\mu} = \sum_{i=1}^{g} p_i \boldsymbol{\mu}_{\mathbf{i}} \tag{5.14}$$

The marginal variance (1.6) is defined as the following.

$$\boldsymbol{\Sigma} = \sum_{i=1}^{g} \boldsymbol{\Sigma}_{\mathbf{i}} p_i + \sum_{j=1}^{g} \boldsymbol{\mu}_{\mathbf{i}} \boldsymbol{\mu}_{\mathbf{i}}' p_i - \boldsymbol{\mu}\boldsymbol{\mu}' \tag{5.15}$$

Recall the definition of $\mathrm{E}[\mathbf{u_i}]$ from equation (3.4). We already know the first g parameters (the group proportions). The pseudo-first moment parameters are easily calculated as $\mathbf{m_j} = p_j \boldsymbol{\mu}_{\mathbf{j}}$. The reader will recall our first delta method inverted this scaling to get the actual group means.

The pseudo second moments are obtained by using the same scaling trick and the relationship between the variance and first and second moments.

$$\begin{aligned}
\mathrm{Var}[\mathbf{x}] &= \mathrm{E}\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'\right] \\
&= \mathrm{E}\left[(\mathbf{x} - \boldsymbol{\mu})\mathbf{x}'\right] - \mathrm{E}\left[(\mathbf{x} - \boldsymbol{\mu})\boldsymbol{\mu}'\right] \\
&= \mathrm{E}\left[(\mathbf{x} - \boldsymbol{\mu})\mathbf{x}'\right] - \mathrm{E}\left[(\mathbf{x} - \boldsymbol{\mu})\right]\boldsymbol{\mu}' \\
&= \mathrm{E}\left[(\mathbf{x} - \boldsymbol{\mu})\mathbf{x}'\right] - \mathbf{0}\boldsymbol{\mu}' \\
&= \mathrm{E}\left[(\mathbf{x} - \boldsymbol{\mu})\mathbf{x}'\right] \\
&= \mathrm{E}\left[\mathbf{x}\mathbf{x}' - \boldsymbol{\mu}\mathbf{x}'\right] \\
&= \mathrm{E}\left[\mathbf{x}\mathbf{x}'\right] - \mathrm{E}\left[\boldsymbol{\mu}\mathbf{x}'\right] \\
&= \mathrm{E}\left[\mathbf{x}\mathbf{x}'\right] - \boldsymbol{\mu}\,\mathrm{E}\left[\mathbf{x}'\right] \\
&= \mathrm{E}\left[\mathbf{x}\mathbf{x}'\right] - \boldsymbol{\mu}\boldsymbol{\mu}'
\end{aligned} \tag{5.16}$$

So we can get the second moment pseudo-parameters in (3.4) by $\mathbf{V_j} = p_j\left(\boldsymbol{\Sigma}_{\mathbf{j}} + \boldsymbol{\mu}_{\mathbf{j}}\boldsymbol{\mu}_{\mathbf{j}}'\right)$. With this we can fully calculate $\mathrm{E}[\mathbf{u_i}]$.

We can already compute much of the variance matrix in $\mathrm{Var}[\mathbf{u_i}]$ (3.5). We still need to derive the $\mathbf{M_j} = \mathrm{E}[1_{ji}\mathbf{x_i}(\mathbf{x_i}' \otimes \mathbf{x_i}')]$, $\mathbf{N_j} = \mathrm{E}\left[1_{ji}(\mathbf{x_i}\mathbf{x_i}') \otimes (\mathbf{x_i}\mathbf{x_i}')\right]$,

$\mathbf{M} = \mathrm{E}[\mathbf{x_i}(\mathbf{x_i'} \otimes \mathbf{x_i'})]$, and $\mathbf{N} = \mathrm{E}\left[(\mathbf{x_i}\mathbf{x_i'}) \otimes (\mathbf{x_i}\mathbf{x_i'})\right]$ parameter matrices. These matrices will utilize higher order powers of the predictors, up to the quartic order. As with the $\mathbf{m_j}$ and $\mathbf{V_j}$ parameters, $\mathbf{M_j} = p_j \mathrm{E}[\mathbf{x_i}\left(\mathbf{x_i'} \otimes \mathbf{x_i'}\right)|y = j]$ and $\mathbf{N_j} = p_j \mathrm{E}\left[(\mathbf{x_i}\mathbf{x_i'}) \otimes (\mathbf{x_i}\mathbf{x_i'})|y = j\right]$.

Let's examine $\mathbf{M_j}$. Conditional on $y = j$, $\mathbf{x_i}$ is a k variate mixture of (or single) multivariate normal random vectors. We will examine what happens when we expand the cubic form in $\mathbf{M_j}$.

$$
\begin{aligned}
&\mathbf{x_i}(\mathbf{x_i'} \otimes \mathbf{x_i'}) = \\
&\begin{bmatrix} x_{i1} \\ \vdots \\ x_{ik} \end{bmatrix} \left(\begin{bmatrix} x_{i1} & \cdots & x_{ik} \end{bmatrix} \otimes \begin{bmatrix} x_{i1} & \cdots & x_{ik} \end{bmatrix} \right) = \\
&\begin{bmatrix} x_{i1} \\ \vdots \\ x_{ik} \end{bmatrix} \begin{bmatrix} x_{i1}x_{i1} & \cdots & x_{i1}x_{ik} & \cdots & x_{ik}x_{i1} & \cdots & x_{ik}x_{ik} \end{bmatrix} = \\
&\begin{bmatrix} x_{i1}x_{i1}x_{i1} & \cdots & x_{i1}x_{i1}x_{ik} & \cdots & x_{i1}x_{ik}x_{i1} & \cdots & x_{i1}x_{ik}x_{ik} \\ \vdots & & \vdots & & \vdots & & \vdots \\ x_{ik}x_{i1}x_{i1} & \cdots & x_{ik}x_{i1}x_{ik} & \cdots & x_{ik}x_{ik}x_{i1} & \cdots & x_{ik}x_{ik}x_{ik} \end{bmatrix}
\end{aligned}
\tag{5.17}
$$

Recall how we calculated the marginal mean $\boldsymbol{\mu}$. We can use the conditioning rule $\mathrm{E}[\mathbf{X}] = \mathrm{E}\left[\mathrm{E}[\mathbf{X}|y]\right]$ to write $\mathbf{M_j}$ as a matrix of linear combinations of multivariate normal moments.

We have a similar situation for $\mathbf{N_j}$. First we examine the quartic form within it.

$$
\begin{aligned}
&(\mathbf{x_i x_i'}) \otimes (\mathbf{x_i x_i'}) = \\
&\left(\begin{bmatrix} x_{i1} \\ \vdots \\ x_{ik} \end{bmatrix} \begin{bmatrix} x_{i1} & \cdots & x_{ik} \end{bmatrix} \right) \otimes \left(\begin{bmatrix} x_{i1} \\ \vdots \\ x_{ik} \end{bmatrix} \begin{bmatrix} x_{i1} & \cdots & x_{ik} \end{bmatrix} \right) = \\
&\begin{bmatrix} x_{i1}x_{i1} & \cdots & x_{i1}x_{ik} \\ \vdots & \ddots & \vdots \\ x_{ik}x_{i1} & \cdots & x_{ik}x_{ik} \end{bmatrix} \otimes \begin{bmatrix} x_{i1}x_{i1} & \cdots & x_{i1}x_{ik} \\ \vdots & \ddots & \vdots \\ x_{ik}x_{i1} & \cdots & x_{ik}x_{ik} \end{bmatrix} = \\
&\begin{bmatrix} x_{i1}x_{i1} \begin{bmatrix} x_{i1}x_{i1} & \cdots & x_{i1}x_{ik} \\ \vdots & \ddots & \vdots \\ x_{ik}x_{i1} & \cdots & x_{ik}x_{ik} \end{bmatrix} & \cdots & x_{i1}x_{ik} \begin{bmatrix} x_{i1}x_{i1} & \cdots & x_{i1}x_{ik} \\ \vdots & \ddots & \vdots \\ x_{ik}x_{i1} & \cdots & x_{ik}x_{i1} \end{bmatrix} \\ \vdots & \ddots & \vdots \\ x_{ik}x_{i1} \begin{bmatrix} x_{i1}x_{i1} & \cdots & x_{i1}x_{ik} \\ \vdots & \ddots & \vdots \\ x_{ik}x_{i1} & \cdots & x_{ik}x_{ik} \end{bmatrix} & \cdots & x_{ik}x_{ik} \begin{bmatrix} x_{i1}x_{i1} & \cdots & x_{i1}x_{ik} \\ \vdots & \ddots & \vdots \\ x_{ik}x_{i1} & \cdots & x_{ik}x_{ik} \end{bmatrix} \end{bmatrix} = \qquad (5.18) \\
&\begin{bmatrix} \begin{bmatrix} x_{i1}x_{i1}x_{i1}x_{i1} & \cdots & x_{i1}x_{i1}x_{i1}x_{ik} \\ \vdots & \ddots & \vdots \\ x_{i1}x_{i1}x_{ik}x_{i1} & \cdots & x_{i1}x_{i1}x_{ik}x_{ik} \end{bmatrix} & \cdots & \begin{bmatrix} x_{i1}x_{ik}x_{i1}x_{i1} & \cdots & x_{i1}x_{ik}x_{i1}x_{ik} \\ \vdots & \ddots & \vdots \\ x_{i1}x_{ik}x_{ik}x_{i1} & \cdots & x_{i1}x_{ik}x_{ik}x_{ik} \end{bmatrix} \\ \vdots & \ddots & \vdots \\ \begin{bmatrix} x_{ik}x_{i1}x_{i1}x_{i1} & \cdots & x_{ik}x_{i1}x_{i1}x_{ik} \\ \vdots & \ddots & \vdots \\ x_{ik}x_{i1}x_{ik}x_{i1} & \cdots & x_{ik}x_{i1}x_{ik}x_{ik} \end{bmatrix} & \cdots & \begin{bmatrix} x_{ik}x_{ik}x_{i1}x_{i1} & \cdots & x_{ik}x_{ik}x_{i1}x_{ik} \\ \vdots & \ddots & \vdots \\ x_{ik}x_{ik}x_{ik}x_{i1} & \cdots & x_{ik}x_{ik}x_{ik}x_{ik} \end{bmatrix} \end{bmatrix}
\end{aligned}
$$

Taking the expectation of this will yield a matrix of fourth order product moments of variables that are mixtures of (or single) multivariate normal.

Look at the following example. Here we let $\mu_j^{m\_kk1}$ be the third order product

moment of between variables x_k,x_k, and x_1 for subpopulation m within group j. If we know the mean vector $\boldsymbol{\mu}_{\mathbf{j}}^{\mathbf{m}}$ and the covariance matrix $\boldsymbol{\Sigma}_{\mathbf{j}}^{\mathbf{m}}$ for subpopulation m within group j, then we can calculate this product moment.

$$\begin{aligned}
&\mathrm{E}\left[x_{ik}x_{ik}x_{i1}|y=j\right] &&= \\
&\mathrm{E}\left[\mathrm{E}\left[x_{ik}x_{ik}x_{i1}\middle|\text{group subpop. is } m\right]\middle|y=j\right] &&= \\
&\mathrm{E}\left[\textstyle\sum_{m=1}^{g_y}\mathrm{I}\left[\text{group subpop. is } m\right]\mu_y^{m_kk1}\middle|y=j\right] &&= \\
&\mathrm{E}\left[\textstyle\sum_{m=1}^{g_j}\mathrm{I}\left[\text{group subpop. is } m\right]\mu_j^{m_kk1}\right] &&= \\
&\textstyle\sum_{m=1}^{g_j}p_j^m\mu_j^{m_kk1}
\end{aligned} \tag{5.19}$$

We can calculate each of the components of (5.18) and (5.17) in a similar manner to that of (5.19).

The situation is a bit more complicated for the matrices $\mathbf{M}$ and $\mathbf{N}$. There is no conditioning in either expression. So the marginals are taken over two levels of mixture: the group, and then the subpopulation within the group (unless the group only has a single multivariate normal population). Calculation is still straightforward however, we just marginalize twice.

We have discussed and provided some demonstration how we can calculate the $\mathrm{E}[\mathbf{u_i}]$ and $\mathrm{Var}[\mathbf{u_i}]$ from (3.22). Through the invocation of several functions this will then give us the values of $\boldsymbol{\Delta}$ and $\mathbf{K}$, and this will allow us to study the theory null distribution (4.27). Our only requirements are that the group proportions p_i, group subpopulation proportions p_i^j, group subpopulation means $\boldsymbol{\mu}_{\mathbf{i}}^{\mathbf{j}}$, and group subpopulation variances $\boldsymbol{\Sigma}_{\mathbf{i}}^{\mathbf{j}}$ be specified. We allow subpopulations within the groups for extra flexibility, but for simplicity we may only use one population within each group.

We can calculate the multivariate normal moments (up to fourth order) for the estimation of each quantity in $\mathrm{E}[\mathbf{u_i}]$ and $\mathrm{Var}[\mathbf{u_i}]$ based solely on our specified pa-

rameters, since the multivariate normal distribution is fully defined by its mean and variance. However, the link between the first and second moments that the mean and variance define and the higher order moments is not elementary. We will define this link, i.e. show how to calculate the higher order moments given the first and second moments in the appendix.

C. Example

Now we will provide examples that demonstrate the use of the concepts we have discussed so far. We examine 2 different situations.

In the first example, we use the statistical programming language R to generate data from a hypothetical 2 group population of 6-variate multivariate normals. This is the same situation we encountered in our initial example in section 1.1. There are four differences between the groups. The mean vectors differ. All of the variance parameters are different between the two groups (yielding 1 difference since we stack the variances). One of the covariance parameters is different, that between x_5 and x_6. As discussed earlier, for reasons of symmetry we find a difference between the covariance matrix parameters at index (5,6) and (6,5), yielding 2 differences.

For sample sizes 100, 1000, and 5000 we show how the approximate empirical (4.30) and empirical (4.28) tests are used to estimate the dimensionality of the SMVCIR space. We will show the results of the iterative tests (4.1) for each of the three separate samples. For the the largest sample, where $n = 5000$, we will show how SMVCIR transforms the coordinates and demonstrates the differences between the two groups.

We will also use another situation to show how the theory (4.27) null reference distribution matches with the distribution of our test statistic. In this situation we

have three groups and four differences for a 5 variable population. There is one mean difference, one variance difference, and two covariance differences.

We merely want to demonstrate how our developed concepts work in this chapter. Further simulation examining the approximate empirical (4.30), empirical (4.28), and theory (4.27) distribution results will be performed in the next chapter. Our claims about the three will be rigorously corroborated there.

1. 2 Groups, 6 Predictors: Approximate Empirical and Empirical

We observe 2 multivariate normal populations with these parameters.

$$\begin{aligned} p_1 &= .25 \\ p_2 &= .75 \\ \boldsymbol{\mu_1'} &= \begin{bmatrix} 17.5 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}' \\ \boldsymbol{\mu_2'} &= \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \end{bmatrix}' \end{aligned} \tag{5.20}$$

$$\begin{aligned} \boldsymbol{\Sigma_1} &= \mathbf{I_6} \\ \boldsymbol{\Sigma_2} &= \begin{bmatrix} 4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 9 & 0 & 0 & 0 & 0 \\ 0 & 0 & 16 & 0 & 0 & 0 \\ 0 & 0 & 0 & 25 & 0 & 0 \\ 0 & 0 & 0 & 0 & 36 & 20 \\ 0 & 0 & 0 & 0 & 20 & 49 \end{bmatrix} \end{aligned} \tag{5.21}$$

So we expect to find 4 difference dimensions when we use SMVCIR. The means are clearly different for the first variable x_1. The variance of all variables $x_1, \ldots, x_6$ is different across the two groups. The covariance between x_5 and x_6 is clearly different as well. The mean difference should contribute one unique difference to the SMVCIR space. All six of the elemental variance differences will contribute in total one unique variance difference to the SMVCIR space (recall that variances are stacked together). We will see two dimensions for the covariance difference between x_5 and x_6. Recall

our explanation in section 1.1 for the covariance difference in this same example situation. We simulate 100 draws from the population using hierarchical sampling and then perform SMVCIR on the resulting sample. In this hierarchical sampling we sample each observation by first drawing a group. We choose group 1 with probability .25 and group 2 with probability .75. Then we sample from the multivariate normal population of that group using the function mvrnorm from the MASS package in R (Venables and Ripley (2002)).

We have written an R program that performs the SMVCIR algorithm and performs the dimensionality tests. Invocation of this program on the size 100 sample yields the following results in Table 1 . We have rounded the p-value and test statistic values to the 1000th decimal place for simplicity.

Table 1. Ex. 5.C.1, Tests for d, $n = 100$

Test	$\hat{\Lambda}_i$	Empirical	Approx. Empirical
$H_0 : d = 0$ $H_1 : d > 0$	235.039	0	0
$H_0 : d = 1$ $H_1 : d > 1$	117.303	0	0
$H_0 : d = 2$ $H_1 : d > 2$	24.115	0	0
$H_0 : d = 3$ $H_1 : d > 3$	12.634	0	0
$H_0 : d = 4$ $H_1 : d > 4$	3.750	0.007	.003
$H_0 : d = 5$ $H_1 : d > 5$	.082	0.857	.829

So SMVCIR detects 5 dimensions under this sample size. We now perform SMVCIR on the size 1000 sample, recording the results in Table 2.

Table 2. Ex. 5.C.1, Tests for d, $n = 1000$

Test	$\hat{\Lambda}_i$	Empirical	Approx. Empirical
$H_0 : d = 0$ $H_1 : d > 0$	2529.054	0	0
$H_0 : d = 1$ $H_1 : d > 1$	1064.658	0	0
$H_0 : d = 2$ $H_1 : d > 2$	115.248	0	0
$H_0 : d = 3$ $H_1 : d > 3$	39.833	0	0
$H_0 : d = 4$ $H_1 : d > 4$	4.076	0.061	.062
$H_0 : d = 5$ $H_1 : d > 5$	1.676	0.101	.111

Using a significance level of .05 and not worrying about the effect of multiple testing on our type 1 error rates, SMVCIR find 4 dimensions.

This is acceptable, but we will see better performance under a sample size of 5000. We will show the output of the separate stages of the SMVCIR algorithm for this next calculation. Since the output matrix size is so large, we provide this output in its original R format. We round the estimates in each of the span set estimation stages to the 10000th decimal place.

In the first stage, the central limit theorem stage recorded in Table 3, the output does not look much like our parameters because we have not yet created the group moments. Recall equation (3.3). We do see the familiar sample group proportions in the first line of the output. These are repeated in the output of each stage until the group proportions are no longer needed. The statistics that will become sample group means make up the first two columns of the remaining output. The statistics that will become sample group second moments make up the next 12 columns (6 columns

for six variables and two groups). The marginal mean is in the next column (16). The marginal variance occupies the last 6 columns.

Table 3. Output: Central Limit Theorem

1	0.2396	0.7604							
	1	2	3	4	5	6	7	8	9
1	4.1944	0.0314	73.6682	-0.1229	0.238	0.111	0.0597	0.1283	3.0514
2	-0.0071	-0.082	-0.1229	0.2262	-0.0009	-0.004	-0.0049	0.0043	-0.1237
3	0.0129	0.0573	0.238	-0.0009	0.2598	0.0039	-0.0127	-0.0111	-0.0259
4	0.0064	-0.1315	0.111	-0.004	0.0039	0.2322	-0.0113	0.0032	-0.2043
5	0.0025	0.0943	0.0597	-0.0049	-0.0127	-0.0113	0.2242	0.0037	0.259
6	0.0077	0.1052	0.1283	0.0043	-0.0111	0.0032	0.0037	0.2326	0.1154
	10	11	12	13	14	15	16	17	18
1	-0.1237	-0.0259	-0.2043	0.259	0.1154	4.2258	76.7196	-0.2466	0.2121
2	6.864	0.0503	0.0082	-0.559	-0.4642	-0.089	-0.2466	7.0902	0.0494
3	0.0503	12.3843	0.0844	0.0227	0.1591	0.0702	0.2121	0.0494	12.6441
4	0.0082	0.0844	19.0061	-0.3936	-0.772	-0.1251	-0.0933	0.0042	0.0883
5	-0.559	0.0227	-0.3936	26.8518	15.1704	0.0968	0.3187	-0.5638	0.01
6	-0.4642	0.1591	-0.772	15.1704	37.847	0.1129	0.2437	-0.4599	0.1481
	19	20	21						
1	-0.0933	0.3187	0.2437						
2	0.0042	-0.5638	-0.4599						
3	0.0883	0.01	0.1481						
4	19.2383	-0.4049	-0.7687						
5	-0.4049	27.076	15.1742						
6	-0.7687	15.1742	38.0796						

Now in the next stage, our first delta method invocation recorded in Table 4, we can see the group mean estimates (first two columns). The group second moments are given next (columns 3-8 for group 1, columns 9-14 for group 2). Since the mean of group 2 is zero, the second moments for that group are understandably close to the variances and covariances. For group 1, note how large the second moment estimate is for variable 1 (row 1 and column 3). This size is due solely to the non-zero mean in group 1 for variable 1 (estimated in row 1 and column 1). The marginal first moment (column 15) and the second moments in the last 6 columns remain the same.

Table 4. Output: Group Moments, f_1

1	0.2396	0.7604							
	1	2	3	4	5	6	7	8	9
1	17.5056	0.0413	307.4634	-0.5127	0.9932	0.4632	0.2491	0.5353	4.0129
2	-0.0296	-0.1078	-0.5127	0.9442	-0.0036	-0.0167	-0.0203	0.0178	-0.1627
3	0.0537	0.0754	0.9932	-0.0036	1.0843	0.0163	-0.0529	-0.0462	-0.034
4	0.0266	-0.1729	0.4632	-0.0167	0.0163	0.9692	-0.0472	0.0135	-0.2686
5	0.0105	0.124	0.2491	-0.0203	-0.0529	-0.0472	0.9357	0.0155	0.3406
6	0.032	0.1384	0.5353	0.0178	-0.0462	0.0135	0.0155	0.9706	0.1518
	10	11	12	13	14	15	16	17	18
1	-0.1627	-0.034	-0.2686	0.3406	0.1518	4.2258	76.7196	-0.2466	0.2121
2	9.0268	0.0662	0.0107	-0.7351	-0.6104	-0.089	-0.2466	7.0902	0.0494
3	0.0662	16.2866	0.1111	0.0298	0.2093	0.0702	0.2121	0.0494	12.6441
4	0.0107	0.1111	24.9949	-0.5176	-1.0152	-0.1251	-0.0933	0.0042	0.0883
5	-0.7351	0.0298	-0.5176	35.3127	19.9506	0.0968	0.3187	-0.5638	0.01
6	-0.6104	0.2093	-1.0152	19.9506	49.7725	0.1129	0.2437	-0.4599	0.1481
	19	20	21						
1	-0.0933	0.3187	0.2437						
2	0.0042	-0.5638	-0.4599						
3	0.0883	0.01	0.1481						
4	19.2383	-0.4049	-0.7687						
5	-0.4049	27.076	15.1742						
6	-0.7687	15.1742	38.0796						

In the next stage, our group second moments are turned into variances. The output is recorded in recorded in Table 5. We see that the high estimate value in row 1 and column 3 has been changed to a value closer to the variance of group 1 for variable 1. The marginal second moment estimates in the last six columns have become marginal variance/covariance estimates.

Table 5. Output: Variances, f_2

1	0.2396	0.7604							
	1	2	3	4	5	6	7	8	9
1	17.5056	0.0413	1.0159	0.0055	0.0525	-0.0019	0.0654	-0.0246	4.0112
2	-0.0296	-0.1078	0.0055	0.9433	-0.0021	-0.0159	-0.02	0.0187	-0.1583
3	0.0537	0.0754	0.0525	-0.0021	1.0814	0.0148	-0.0534	-0.0479	-0.0371
4	0.0266	-0.1729	-0.0019	-0.0159	0.0148	0.9685	-0.0475	0.0127	-0.2615
5	0.0105	0.124	0.0654	-0.02	-0.0534	-0.0475	0.9356	0.0152	0.3355
6	0.032	0.1384	-0.0246	0.0187	-0.0479	0.0127	0.0152	0.9696	0.1461
	10	11	12	13	14	15	16	17	18
1	-0.1583	-0.0371	-0.2615	0.3355	0.1461	4.2258	58.8625	0.1297	-0.0845
2	9.0152	0.0743	-0.0079	-0.7217	-0.5955	-0.089	0.1297	7.0822	0.0557
3	0.0743	16.2809	0.1241	0.0205	0.1989	0.0702	-0.0845	0.0557	12.6392
4	-0.0079	0.1241	24.965	-0.4962	-0.9913	-0.1251	0.4354	-0.007	0.0971
5	-0.7217	0.0205	-0.4962	35.2973	19.9334	0.0968	-0.0903	-0.5552	0.0032
6	-0.5955	0.1989	-0.9913	19.9334	49.7533	0.1129	-0.2333	-0.4499	0.1402
	19	20	21						
1	0.4354	-0.0903	-0.2333						
2	-0.007	-0.5552	-0.4499						
3	0.0971	0.0032	0.1402						
4	19.2227	-0.3928	-0.7546						
5	-0.3928	27.0666	15.1632						
6	-0.7546	15.1632	38.0668						

In the third delta method stage, the data is standardized. We record this output in Table 6. The marginal mean and variance information is dropped from the output. Note how small the variance in group 1 (columns 3-8) has become. Our marginal variance is greatly affected by the high variance in group 2 and the mean of group 1.

Table 6. Output: Standardization, f_3

1	0.2396	0.7604							
	1	2	3	4	5	6	7	8	9
1	1.7309	-0.5454	0.0173	0.0003	0.0019	-0.0001	0.0016	-0.0005	0.0681
2	0.0223	-0.007	0.0003	0.1332	-0.0002	-0.0014	-0.0014	0.0011	-0.0078
3	-0.0046	0.0015	0.0019	-0.0002	0.0856	0.001	-0.0029	-0.0022	-0.0014
4	0.0346	-0.0109	-0.0001	-0.0014	0.001	0.0504	-0.0021	0.0005	-0.0078
5	-0.0166	0.0052	0.0016	-0.0014	-0.0029	-0.0021	0.0346	0.0005	0.0084
6	-0.0131	0.0041	-0.0005	0.0011	-0.0022	0.0005	0.0005	0.0255	0.0031
	10	11	12	13	14				
1	-0.0078	-0.0014	-0.0078	0.0084	0.0031				
2	1.2729	0.0079	-0.0007	-0.0521	-0.0363				
3	0.0079	1.2881	0.008	0.0011	0.0091				
4	-0.0007	0.008	1.2987	-0.0218	-0.0366				
5	-0.0521	0.0011	-0.0218	1.3041	0.621				
6	-0.0363	0.0091	-0.0366	0.621	1.307				

In the fourth stage, we center the variances using the pooled variances. The output is recorded in Table 7.

Table 7. Output: Centered Variances, f_4

1	0.2396	0.7604							
	1	2	3	4	5	6	7	8	9
1	1.7309	-0.5454	-0.0387	0.0061	0.0025	0.0059	-0.0051	-0.0027	0.0122
2	0.0223	-0.007	0.0061	-0.8667	-0.0061	-0.0005	0.0385	0.0284	-0.0019
3	-0.0046	0.0015	0.0025	-0.0061	-0.9144	-0.0053	-0.003	-0.0086	-0.0008
4	0.0346	-0.0109	0.0059	-0.0005	-0.0053	-0.9492	0.015	0.0282	-0.0018
5	-0.0166	0.0052	-0.0051	0.0385	-0.003	0.015	-0.9653	-0.4718	0.0016
6	-0.0131	0.0041	-0.0027	0.0284	-0.0086	0.0282	-0.4718	-0.9745	0.0009
	10	11	12	13	14				
1	-0.0019	-0.0008	-0.0018	0.0016	0.0009				
2	0.2731	0.0019	0.0002	-0.0121	-0.009				
3	0.0019	0.2881	0.0017	0.001	0.0027				
4	0.0002	0.0017	0.2991	-0.0047	-0.0089				
5	-0.0121	0.001	-0.0047	0.3042	0.1487				
6	-0.009	0.0027	-0.0089	0.1487	0.3071				

In stage 5 we weight the group estimates. The group proportions are no longer needed and dropped from the output. The results are recorded in Table 8.

Table 8. Output: Group Proportion Weighted, f_5

	1	2	3	4	5	6
1	0.8473	-0.4756	-0.0189	0.003	0.0012	0.0029
2	0.0109	-0.0061	0.003	-0.4242	-0.003	-0.0003
3	-0.0023	0.0013	0.0012	-0.003	-0.4476	-0.0026
4	0.0169	-0.0095	0.0029	-0.0003	-0.0026	-0.4646
5	-0.0081	0.0046	-0.0025	0.0189	-0.0015	0.0073
6	-0.0064	0.0036	-0.0013	0.0139	-0.0042	0.0138
	10	11	12	13	14	
1	-0.0017	-0.0007	-0.0016	0.0014	0.0008	
2	0.2381	0.0017	0.0001	-0.0106	-0.0078	
3	0.0017	0.2513	0.0015	0.0008	0.0024	
4	0.0001	0.0015	0.2608	-0.0041	-0.0078	
5	-0.0106	0.0008	-0.0041	0.2652	0.1296	
6	-0.0078	0.0024	-0.0078	0.1296	0.2678	

The final spanning set estimate is then created. We record the estimate in Table 9. Note how zeros have replaced the variance positions in the weighted group covariances, and how the variances that were there are stacked at the end in the final two columns.

Table 9. Output: Estimated Spanset from Simulation Functions

	1	2	3	4	5	6	7	8	9
1	0.8473	-0.4756	0	0.003	0.0012	0.0029	-0.0025	-0.0013	0
2	0.0109	-0.0061	0.003	0	-0.003	-0.0003	0.0189	0.0139	-0.0017
3	-0.0023	0.0013	0.0012	-0.003	0	-0.0026	-0.0015	-0.0042	-0.0007
4	0.0169	-0.0095	0.0029	-0.0003	-0.0026	0	0.0073	0.0138	-0.0016
5	-0.0081	0.0046	-0.0025	0.0189	-0.0015	0.0073	0	-0.231	0.0014
6	-0.0064	0.0036	-0.0013	0.0139	-0.0042	0.0138	-0.231	0	0.0008
	10	11	12	13	14	15	16		
1	-0.0017	-0.0007	-0.0016	0.0014	0.0008	-0.0189	0.0106		
2	0	0.0017	0.0001	-0.0106	-0.0078	-0.4242	0.2381		
3	0.0017	0	0.0015	0.0008	0.0024	-0.4476	0.2513		
4	0.0001	0.0015	0	-0.0041	-0.0078	-0.4646	0.2608		
5	-0.0106	0.0008	-0.0041	0	0.1296	-0.4725	0.2652		
6	-0.0078	0.0024	-0.0078	0.1296	0	-0.477	0.2678		

We just saw the spanning set used to generate the reference distribution in the approximate empirical (4.30) and empirical (4.28), Table 9 "Estimated Spanset from Simulation Functions". The actual computation of the SMVCIR spanning set and test statistic in practice use bias corrected variance estimates. However, the estimate produced for these reference distributions, Table 9 "Estimated Spanset from Simulation Functions", which includes the separate stages we have shown, uses the biased statistics for simplicity. As we discussed in Chapter III, the biased and unbiased estimates used in SMVCIR are both consistent and asymptotically equivalent to each other.

The spanning set that is used to compute the test statistic is the following, as recorded in Table 10

Table 10. Output: Estimated Spanset from Original Smvcir1 Function

	1	2	3	4	5	6	7	8	9
1	0.8472	-0.4756	0	0.003	0.0012	0.0029	-0.0025	-0.0013	0
2	0.0109	-0.0061	0.003	0	-0.003	-0.0003	0.0189	0.0139	-0.0017
3	-0.0023	0.0013	0.0012	-0.003	0	-0.0026	-0.0015	-0.0042	-0.0007
4	0.0169	-0.0095	0.0029	-0.0003	-0.0026	0	0.0073	0.0138	-0.0016
5	-0.0081	0.0046	-0.0025	0.0189	-0.0015	0.0073	0	-0.231	0.0014
6	-0.0064	0.0036	-0.0013	0.0139	-0.0042	0.0138	-0.231	0	0.0008

	10	11	12	13	14	15	16
1	-0.0017	-0.0007	-0.0016	0.0014	0.0008	-0.0189	0.0106
2	0	0.0017	0.0001	-0.0106	-0.0078	-0.4242	0.2381
3	0.0017	0	0.0015	0.0008	0.0024	-0.4476	0.2513
4	0.0001	0.0015	0	-0.0041	-0.0078	-0.4647	0.2608
5	-0.0106	0.0008	-0.0041	0	0.1297	-0.4726	0.2653
6	-0.0078	0.0024	-0.0078	0.1297	0	-0.477	0.2678

It is very similar to the previous spanning set, Table 9 “Estimated Spanset from Simulation Functions”. The kernel of the estimate Table 10 “Estimated Spanset from Original Smvcir1 Function” has the eigenvalues recorded in Table 11. These are the squared singular values of the Table 10 “Estimated Spanset from Original Smvcir1 Function” matrix.

Table 11. Output: Estimated Kernel from Original Smvcir1 Function Eigenvalues

```
1 1.404967e+00 9.430177e-01 7.025956e-02 4.631225e-02 4.562671e-05
6 5.678140e-08
```

The kernel of the spanning set used for the reference distributions, Table 9 “Estimated Spanset from Simulation Functions” has the eigenvalues recorded in Table 12.. They are quite close.

Table 12. Output: Estimated Kernel from Simulation Functions Eigenvalues

```
1 1.404873e+00 9.432050e-01 7.025077e-02 4.630770e-02 4.562375e-05
6 5.662135e-08
```

Now we produce the dimensionality test output, recorded in Table 13. The reference distributions here have been determined via estimated singular value decomposition and covariance of the Table 9 “Estimated Spanset from Simulation Functions” estimated. The test statistic is determined using the Table 10 “Estimated Spanset from Original Smvcir1 Function” spanning set estimate.

We see that the sample size has become adequately high so that the dimension is perfectly inferred. Now we can infer that SMVCIR spans only four dimensions. So we can perform the final analysis of our data by visually examining the first four dimensions of the final transformed data.

Before examination of the transformed data, we check the correlation of the

Table 13. Ex. 5.C.1, Tests for d, $n = 5000$

Test	$\hat{\Lambda}_i$	Empirical	Approx. Empirical
$H_0 : d = 0$ $H_1 : d > 0$	12323.0100	0	0
$H_0 : d = 1$ $H_1 : d > 1$	5298.1760	0	0
$H_0 : d = 2$ $H_1 : d > 2$	583.0875	0	0
$H_0 : d = 3$ $H_1 : d > 3$	232.7897	0	0
$H_0 : d = 4$ $H_1 : d > 4$	.2284	0.9554	.9366
$H_0 : d = 5$ $H_1 : d > 5$	.0003	1	.9996

transformed coordinates. The results are recorded in Table 14. We only need to check the first four since we accepted that our dimension is 4. We find little correlation. So

Table 14. Output: SMVCIR D1-D4 Correlation

	D1	D2	D3	D4
D1	1.0000	0.0124	0.0008	0.1838
D2	0.0124	1.0000	-0.0035	0.0289
D3	0.0008	-0.0035	1.0000	0.0057
D4	0.1838	0.0289	0.0057	1.0000

we can examine the plots without worrying about the problems caused by marginal linear dependency. Here they are in Figure 13. The first group is black and the second is red.

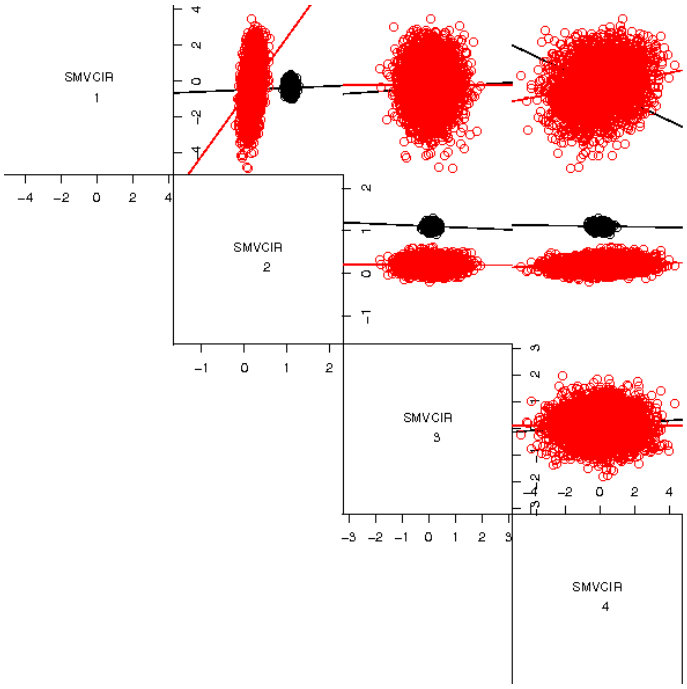

Figure 13. SMVCIR Coordinates, Ex. 5.C.1

The first dimension clearly separates the groups in variance. The second also clearly separates the groups on location. The last two dimensions separate on covariance, though the extent of the separation is much more mild. This is not surprising. The third and fourth theoretical eigenvalues were significantly lower than the first two.

Now we'll look at a 3D plot of the first three dimensions in Figure 14. We use the rgl package (Adler and Murdoch (2009)). All the group differences are noticeable here. The covariance difference is the weakest trend in the examination of the first three dimensions, but it is very evident in the second plot where we examine dimensions 1, 2, and 4.

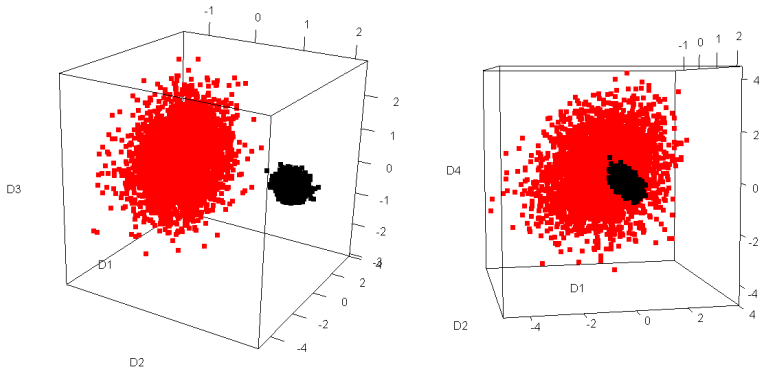

Figure 14. 3D SMVCIR Coordinates, Ex. 5.C.1

Table 15 records the the eigenvectors of the first four dimensions

Table 15. Output: SMVCIR Eigenvectors

	1	2	3	4
1	-0.0662	0.9974	-0.0021	-0.0230
2	-0.4034	-0.0153	-0.0146	0.4340
3	-0.4285	-0.0329	-0.0082	0.3111
4	-0.4435	-0.0111	0.0202	0.4203
5	-0.4734	-0.0441	-0.7096	-0.5167
6	-0.4780	-0.0418	0.7041	-0.5203

The first dimension, found in column 1 represents variance. It involves all the variables and assigns them all the same sign. Most are given an appreciable magnitude as well. This makes sense because the variance is uniformly larger in group 2 than group 1, so the differences are unidirectional.

The second dimension represents the location difference in x_1. The coefficients of the eigenvector are very small except for the first position. The coefficient in this position is also positive, which differs from the coefficients used for the first dimension. The mean of group 1 exceeds the mean of group 2, while the variances of group 2 exceed those of group 1. So the pattern of signs is consistent.

In the third eigenvector, all of the coefficients are very small except for the last two. This puts most of the discrimination strength of this dimension in variables x_5 and x_6 and corroborates our interpretation that this a covariance dimension involving those two variables.

The fourth dimension is more nebulous. It appears to be a contrast between the second through fifth variables with the rest. The fifth and six variables do have the largest magnitude and both point in the same direction. Each SMVCIR dimension does not necessarily coincide with a single variance/covariance difference and may be the result of several. This dimension probably places extra and negative scaling on variables x_5 and x_6 so that they are well contrasted with the positively scaled variables $x_2, \ldots, x_4$, since they have positive covariance and higher variance than those variables. The coefficient on x_1 is negligible in magnitude, so its effect on the dimension is probably rather mild.

Note that these eigenvectors are coefficients for the standardized data in forming the SMVCIR space. Thankfully, as we demonstrated in Chapter II, our scaling standardization method maintains the ordering relationships among the variables between groups.

We can see how SMVCIR succinctly shows us the group differences by comparing Figure 13 with a matrix plot of the original predictors. We produce this matrix plot in Figure 15.

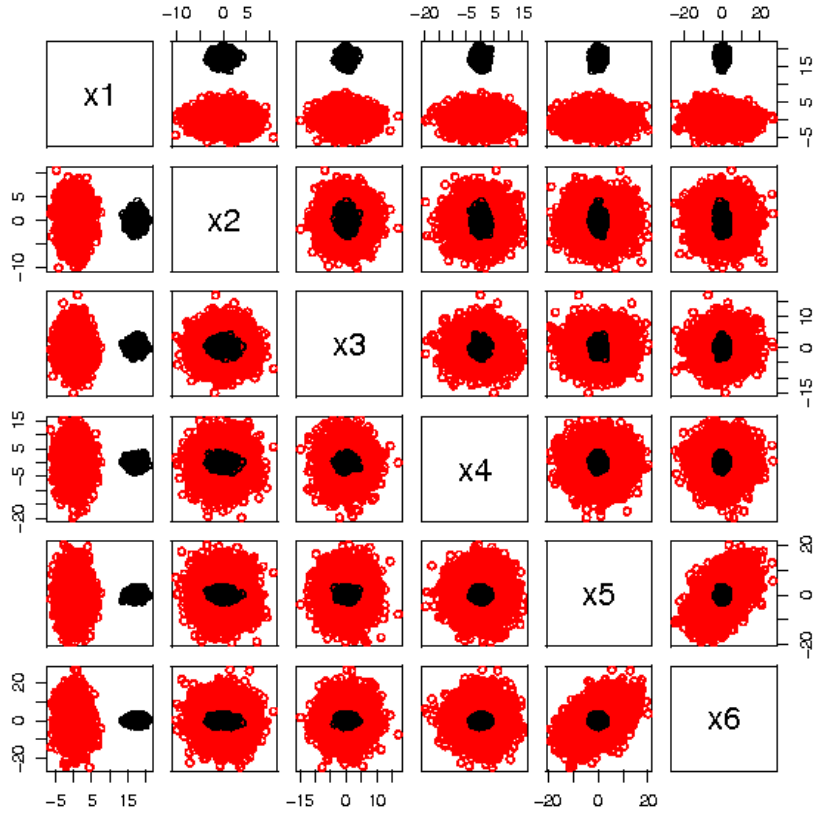

Figure 15. Original Predictor Coordinates, Ex. 5.C.1

2. 3 Groups, 6 Predictors: Theory

In this section we will perform some simulations showing how the theory (4.27) null reference distribution holds for our test statistic $\hat{\Lambda}_i$ under $H_0 : d = i$.

We observe 3 multivariate normal populations with these parameters.

$$
\begin{aligned}
p_1 &= .25 \\
p_2 &= .25 \\
p_3 &= .5 \\
\boldsymbol{\mu}_1' &= \boldsymbol{\mu}_3' = \begin{bmatrix} 2 & .2 & -2 & 3.2 & .6 \end{bmatrix}' \\
\boldsymbol{\mu}_2' &= \begin{bmatrix} .1 & 2 & .3 & .4 & 5 \end{bmatrix}'
\end{aligned}
\tag{5.22}
$$

$$
\begin{aligned}
\boldsymbol{\Sigma}_1 &= \begin{bmatrix} 10 & 6 & 0 & 0 & 0 \\ 6 & 10 & 0 & 0 & 0 \\ 0 & 0 & 10 & 0 & 0 \\ 0 & 0 & 0 & 10 & 0 \\ 0 & 0 & 0 & 0 & 10 \end{bmatrix} \\
\boldsymbol{\Sigma}_2 &= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 10 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 4 \end{bmatrix} \\
\boldsymbol{\Sigma}_3 &= 10 \times \mathbf{I}_5
\end{aligned}
\tag{5.23}
$$

The mean of group 2 is obviously different from that of the other groups. The variances of group 2, except for x_3 are also different from the rest. Group 1 has a different covariance structure than groups 2 or 3, with positive correlation between x_1 and x_2. To reflect this contrast in covariance between group 3 and groups 1 and 2, SMVCIR will put two unique dimensions into the spanning set. The first is a dimension of the standardized difference of column 2 for the groups with zero positions in all but the first row (variance position for x_2 is already zeroed out). The second is a dimension of the standardized difference of column 1 for the groups with zero positions in all but the second row (variance position for x_1 is already zeroed out). So we should expect to see 4 dimensions.

We calculate the theoretical values of the spanning set and the covariance of its

estimate using the tools developed in the previous section. Then we make draws from this theoretical distribution. We compare the distribution estimated by these draws to the distribution with that of the test statistic $\hat{\Lambda}_4$ as estimated over many samples from the population with parameters (5.22)-(5.23). The two should match, as the dimension of the SMVCIR space is 4.

The eigen values for the SMVCIR kernel are the following.

$$\begin{aligned} \lambda_1 &= .7401373 \\ \lambda_2 &= .5996325 \\ \lambda_3 &= .07067002 \\ \lambda_4 &= .02682476 \\ \lambda_5 &= 0 \end{aligned} \tag{5.24}$$

When we use $n = 100$ for our calculation of the test statistic in each sample, we get the following results. A Q-Q plot is rendered in Figure 16. Two kernel density plots are rendered in Figure 17. In the kernel density plots, the red line represents the reference distribution density estimate, while the black represents the test statistic's density estimate. 1000 realizations were drawn from the theoretical and actual distributions to make these plots.

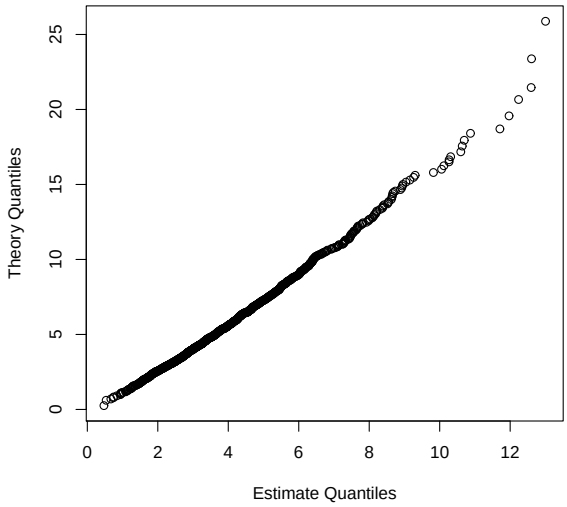

Figure 16. $n = 100$, Q-Q Plot, Ex. 5.C.2

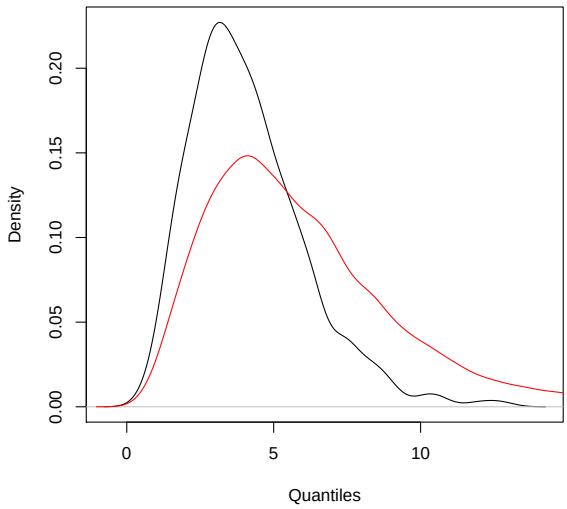

Figure 17. $n = 100$, Kernel Densities, Ex. 5.C.2

At this sample size, the distribution of the test statistic $\hat{\Lambda}_4$ does not match with the theory null reference distribution (4.27). Remember that our results are asymptotic. Looking at the higher sample size, $n = 1000$, we find much better results. We note that outliers in the right tail do not seem to match perfectly between the two distributions. Again 1000 draws were used from each of the theoretical and actual distributions. The Q-Q Plot is rendered in Figure 18. The two kernel densities are rendered in Figure 19.

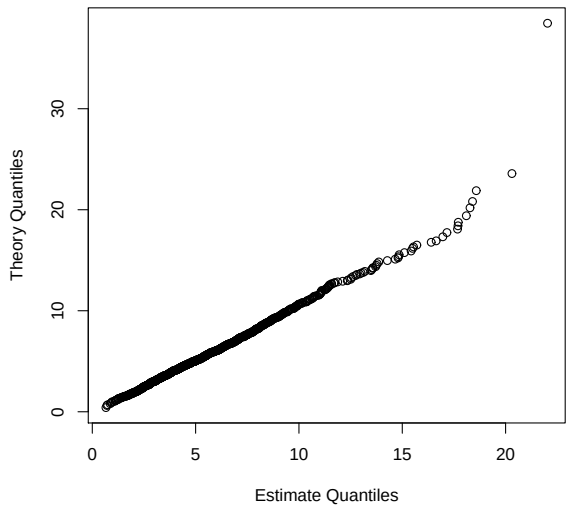

Figure 18. $n = 1000$, Q-Q Plot, Ex. 5.C.2

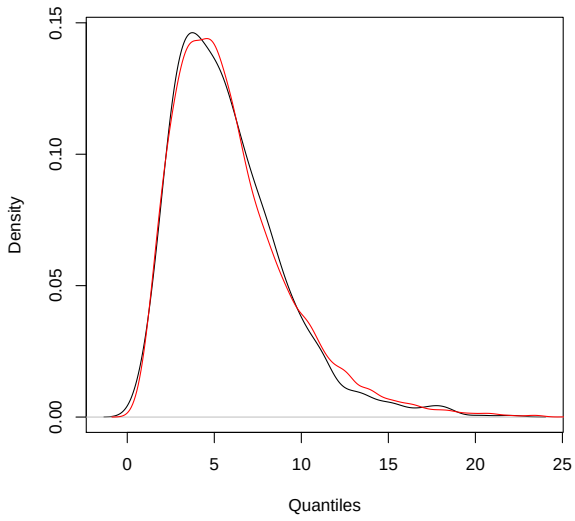

Figure 19. $n = 1000$, Kernel Densities, Ex. 5.C.2

CHAPTER VI

STATISTICAL AND HEURISTIC SIMULATIONS

In this chapter we will rigorously examine our three dimensionality tests: theory (4.27), empirical (4.28), and approximate empirical (4.30) using simulated data from a variety of situations. We will examine the size and power of our SMVCIR dimensionality test using the tools we developed in the last chapter (the calculation tools from Sections A and B of Chapter V and the Appendix). We will study the theory (4.27) reference distribution graphically using kernel density estimates. We demonstrated this method of evaluation in Subsection V.C.2. Multivariate normal, mixture multivariate normal, and other non-normal population settings will be used.

We will also estimate values of d in these settings. We will do so according to the iterated testing scheme (4.1) in many simulated draws from the setting population. This will be a "heuristic" simulation. We will use statistical inference to accept or reject hypotheses as we progress through (4.1), but we will not evaluate the final result's validity with statistical confidence. To do this we would have to examine the testing procedure as a multiple hypothesis test situation. This can be complicated, and will be a direction of future research as we mentioned in the last chapter.

We will be content that our testing scheme works well (4.1) in a setting if it selects the correct value of d reasonably often, and follows a sensible pattern of selection of the other incorrect values (i.e. selects values close to the correct value more often than those far from it).

These informal simulations (4.1) will also allow us to compare the SMVCIR dimensionality test with that of SIR and SAVE. The power and size of the SIR and SAVE tests have been studied extensively elsewhere (for example Yin (2005) and Li (1991)). We will estimate the value of d in SIR and SAVE using a similar testing

scheme to (4.1) for each of the simulated samples of a setting. The tests used for SIR and SAVE are the defaults used in the dr package (Weisberg (2009)). They use the Bentler-Xie approximation (Bentler and Xie (2000)) and do not assume that the data comes from a normal distribution.

We will compare the chosen values of d across SMVCIR, SIR, and SAVE. Our informal simulations will let us make broad comparisons between the three methods. Only broad comparisons are appropriate, as the details of the spanning set and kernel definitions for each method are significantly different.

In the first section, we will perform these informal simulations for (4.1) and evaluate the empirical (4.28) and approximate empirical (4.30) tests using a family of 3 group and 8 predictor distributions. We will evaluate 6 different mean, variance, and covariance scenarios. The scenarios will be imposed upon 3 separate base distributions. The first is normal. The second is a multivariate Student's T distribution that is parameterized by the given mean and variance parameters and has 10 degrees of freedom (only 2 more than its variate dimension). Note that the actual variance of the T distribution will be $10/(10-2) = 1.25$ times the given parameters. In determining group differences, this will have a negligible effect. The third is created by standardizing 8 independent exponential(1) variables. Three of these variables are scaled by negative 1, so their direction of skew is reversed. We first examine the size and power of the empirical (4.28) and approximate empirical (4.30) tests for these settings. At the end of the section we provide the informal simulations for choosing d in SMVCIR, SIR, and SAVE.

In the second section we will test the theory (4.27) reference distribution. We will use quantile vs. quantile plots (Q-Q Plots) and kernel density estimation to evaluate the theory (4.27) reference distribution under multivariate normal and mixture multivariate normal population settings. We choose the mixture multivariate popu-

lation using univariate mixture normals described by Marron and Wand in Marron and Wand (1992).

A. Empirical and Approximate Empirical

We will explore 6 separate mean, variance, covariance configurations. They yield $d = 0, 1, 2, 4, 5, 6$ under SMVCIR. We will first ascertain the size of our test statistic under each dimension. The R programs used to generate the results in Chapter IV will be used again to create our simulated data and perform the tests. The exact nature of the mean, variance, and covariance configuration for each value of d will be described as we present the size results for that configuration.

1. Test Size

We index our parameters vectors and matrices using two numbers. The first is the true number of dimensions. The second is the group number.

The group proportions are the same for each situation.

$$\begin{aligned} p_1 = p_2 &= .25 \\ p_3 &= .5 \end{aligned} \tag{6.1}$$

The group means for the $d = 0$ case are the following.

$$\boldsymbol{\mu}'_{\mathbf{01}} = \boldsymbol{\mu}'_{\mathbf{02}} = \boldsymbol{\mu}'_{\mathbf{03}} = \begin{bmatrix} 5 & -5 & 3 & 2 & 1 & 6 & 7 & 0 \end{bmatrix}' \tag{6.2}$$

The group variances for the $d = 0$ case are given by the following.

$$\boldsymbol{\Sigma_{01}} = \boldsymbol{\Sigma_{02}} = \boldsymbol{\Sigma_{03}} = \begin{bmatrix} 10 & 0 & 0 & 0 & 0 & 0 & 0 & 6 \\ 0 & 12 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 9 & 7 & -2 & 0 & 0 & 3 \\ 0 & 0 & 7 & 11 & -3 & 0 & 0 & 6 \\ 0 & 0 & -2 & -3 & 10 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 9 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 10 & 0 \\ 6 & 0 & 3 & 6 & 1 & 0 & 0 & 14 \end{bmatrix} \tag{6.3}$$

The eigenvalues of the resulting SMVCIR kernel are the following.

$$\lambda_{01} = \ldots = \lambda_{08} = 0 \tag{6.4}$$

We first apply these parameters to a multivariate normal population. We generate samples of $n = 100, 1000, 5000$ and evaluate $\hat{\Lambda}_0$ on each sample. We may decide to reject or accept the null hypothesis $\mathrm{H}_0 : d = 0$ based on the empirical (4.28) reference distribution, using the calculated $\hat{\Lambda}_0$ and the eigenvalues of the null diagonalized spanning set covariance estimate. For each empirical (4.28) test we compute, we simulate 1000 draws to estimate the empirical (4.28) reference distribution. A critical value to use for our decision rule is formulated using this distribution estimate.

The approximate empirical (4.30) reference distribution may also be used. Here we calculate $\hat{\Lambda}_0$ and the null diagonalized spanning set covariance estimate. Then we compute the correction scaling factor and corrected degrees of freedom according to (4.30). A decision rule is then formulated using the resulting χ^2 distribution.

In each of the simulations we create and test 1000 samples for each given sample size and type of test. We use $\alpha = .05$. Note that this is just the nominal size. We are testing the actual size now. Also, treating the size as a proportion, we calculate an exact 95% confidence interval using the Clopper-Pearson technique (Clopper and Pearson (1935)). This helps us see how the size may range in repeated simulations. We begin with the $d = 0$ test size results for the normal population in Table 16.

Table 16. $d = 0$, Normal

Type	n	$\hat{\alpha}$	Exact 95% CI	
Empirical	100	0.06	0.0461	0.0766
	1000	0.055	0.0417	0.0710
	5000	0.068	0.0532	0.0854
Ap. Empirical	100	0.045	0.0330	0.0598
	1000	0.047	0.0347	0.0620
	5000	0.06	0.0461	0.0766

The results suggest that the distributional convergence of $\hat{\Lambda}_0$ under the null hypothesis happens fairly quickly. The confidence intervals for the size all contain the correct value .05, except for the confidence interval for the approximate empirical test under $n = 5000$. The lower bound of this interval is just a little above .05. Now we examine the $d = 0$ test size results for the T_{10} population in Table 17.

Table 17. $d = 0$, T_{10}

Type	n	$\hat{\alpha}$	Exact 95% CI	
Empirical	100	0.039	0.0279	0.0529
	1000	0.038	0.027	0.0518
	5000	0.038	0.027	0.0518
Ap. Empirical	100	0.054	0.0408	0.0699
	1000	0.033	0.0228	0.0460
	5000	0.053	0.0399	0.0688

We see similar results under the multivariate T distribution with 10 degrees of freedom. Convergence under the null occurs quickly. We do note that the estimated size is uniformly smaller than that of Table 17. Also, as in Table 16, one of the confidence intervals fails to include the correct size .05. The upper bound of the confidence interval for the approximate empirical test under $n = 1000$ misses the correct size by .004. Now we examine the $d = 0$ test size results for the exponential population in Table 18.

Table 18. $d = 0$, Standardized Exponential(1)

Type	n	$\hat{\alpha}$	Exact 95% CI	
Empirical	100	0.064	0.0496	0.081
	1000	0.049	0.0365	0.0643
	5000	0.049	0.0365	0.0643
Ap. Empirical	100	0.058	0.0443	0.0743
	1000	0.057	0.0435	0.0732
	5000	0.042	0.0304	0.0564

In this case, we draw independent standardized exponential(1) random variables. Three of the variables x_1, x_3, x_8 have been scaled by -1 so that the direction of skew is not uniform across all eight predictors. We see that all of the confidence intervals include the correct size, .05.

So we have satisfactory size results for the test of $\mathrm{H}_0 : d = 0$ under all three of the our distributional settings. Together, our distributional settings represent a fairly general group of continuous random variable population contexts. So we feel confident that our test of $\mathrm{H}_0 : d = 0$ will have good size and level accuracy when it is used in practice. This should hold true whether the empirical (4.28) or approximate

empirical (4.30) null reference distributions are used.

The $d = 1$ changes the mean vector of the first group. The changed locations are shown in bold.

$$\boldsymbol{\mu}'_{\mathbf{11}} = \left[\begin{array}{cccccccc} \mathbf{-1} & \mathbf{-1} & \mathbf{0} & 2 & 1 & 6 & \mathbf{0} & 0 \end{array} \right]' \tag{6.5}$$

The other parameters remain the same.

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{12}} &= \boldsymbol{\mu}_{\mathbf{02}} \\ \boldsymbol{\mu}_{\mathbf{13}} &= \boldsymbol{\mu}_{\mathbf{03}} \\ \boldsymbol{\Sigma}_{\mathbf{11}} &= \boldsymbol{\Sigma}_{\mathbf{01}} \\ \boldsymbol{\Sigma}_{\mathbf{12}} &= \boldsymbol{\Sigma}_{\mathbf{02}} \\ \boldsymbol{\Sigma}_{\mathbf{13}} &= \boldsymbol{\Sigma}_{\mathbf{03}} \end{aligned} \tag{6.6}$$

The eigenvalues of the SMVCIR kernel for the $d = 1$ are given next.

$$\begin{aligned} \lambda_{11} &= 1.239707 \\ \lambda_{12} &= \ldots = \lambda_{18} = 0 \end{aligned} \tag{6.7}$$

Table 19. $d = 1$, Normal

Type	n	$\hat{\alpha}$	Exact 95% CI	
Empirical	100	0.07	0.055	0.0876
	1000	0.049	0.0365	0.0643
	5000	0.055	0.0417	0.071
Ap. Empirical	100	0.059	0.0452	0.0754
	1000	0.047	0.0347	0.062
	5000	0.065	0.0505	0.0821

In the normal population context, we find that our estimated sizes recorded in Table 19 for $H_0 : d = 1$ are suitable. The empirical test for $n = 100$ and the approximate empirical test under $n = 5000$ have high size. The size confidence interval range for the latter test nearly includes the correct size, .05. The former exhibits far better size for the $n = 1000, 5000$ cases. Now we examine the $d = 1$ test size results for the T_{10} population in Table 20.

Table 20. $d = 1$, T_{10}

Type	n	$\hat{\alpha}$	Exact 95% CI	
Empirical	100	0.079	0.063	0.0975
	1000	0.046	0.0339	0.0609
	5000	0.04	0.0287	0.0541
Ap. Empirical	100	0.078	0.0621	0.0964
	1000	0.032	0.022	0.0449
	5000	0.06	0.0461	0.0766

The T_{10} context shows suitable size results for $H_0 : d = 1$ as well. Under $n = 100$, the size is unacceptably large. This is fine, as the results for the higher sample sizes are much better. For $n = 1000$, the approximate empirical test size estimate is slightly too small as well. Once $n = 5000$, both tests have confidence intervals which include the correct .05 test size.

At $n = 100$, the tests in the standardized exponential(1) context have clearly not converged, as shown in Table 21. The size estimates are unacceptably large for both types of tests. For the higher sample sizes convergence has clearly been attained, and the confidence intervals hold the correct size .05 deep within their range.

We have satisfactory size results for the test of $H_0 : d = 1$ in each of our three

Table 21. $d = 1$, Standardized Exponential(1)

Type	n	$\hat{\alpha}$	Exact 95% CI	
Empirical	100	0.115	0.0959	0.1364
	1000	0.043	0.0313	0.0575
	5000	0.052	0.0391	0.0676
Ap. Empirical	100	0.146	0.1247	0.1694
	1000	0.048	0.0356	0.0631
	5000	0.045	0.033	0.0598

populations contexts. Given their breadth, we feel confident that our test of $\mathrm{H}_0 : d = 1$ will have good size and level accuracy when it is used in practice. This should hold true whether the empirical (4.28) or approximate empirical (4.30) null reference distributions are used.

For $d = 2$, we take the $d = 1$ case and change the variances of group 3. This is the new variance matrix $\mathbf{\Sigma_{23}}$. As before, differing entries from $\mathbf{\Sigma_{13}}$ are printed in boldface.

$$\mathbf{\Sigma_{23}} = \begin{bmatrix} \mathbf{20} & 0 & 0 & 0 & 0 & 0 & 0 & 6 \\ 0 & \mathbf{22} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 9 & 7 & -2 & 0 & 0 & 3 \\ 0 & 0 & 7 & 11 & -3 & 0 & 0 & 6 \\ 0 & 0 & -2 & -3 & 10 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{19} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{30} & 0 \\ 6 & 0 & 3 & 6 & 1 & 0 & 0 & \mathbf{24} \end{bmatrix} \tag{6.8}$$

The other parameters remain the same.

$$\begin{aligned} \boldsymbol{\mu_{21}} &= \boldsymbol{\mu_{11}} \\ \boldsymbol{\mu_{22}} &= \boldsymbol{\mu_{12}} \\ \boldsymbol{\mu_{23}} &= \boldsymbol{\mu_{13}} \\ \boldsymbol{\Sigma_{21}} &= \boldsymbol{\Sigma_{11}} \\ \boldsymbol{\Sigma_{22}} &= \boldsymbol{\Sigma_{12}} \end{aligned} \tag{6.9}$$

The eigenvalues for the resulting SMVCIR kernel are the following.

$$\begin{aligned} &\lambda_{21} = 1.0179 \\ &\lambda_{22} = 0.3447 \\ &\lambda_{23} = \ldots = \lambda_{28} = 0 \end{aligned} \tag{6.10}$$

In the normal population context (recorded in Table 22), we find that our estimated sizes for $\mathrm{H}_0 : d = 2$ are suitable at all the used sample sizes. This suggests that $\hat{\Lambda}_2$ converges in distribution quickly in this case.

The size results for the T_{10} context are recorded in Table 23. The size for $n = 100$ is too high. However, once we increase n the size becomes appropriate. The size for $n = 1000$ of the approximate empirical test is slightly too low, but the confidence interval misses the correct size .05 by less than .002.

Table 22. $d = 2$, Normal

Type	n	$\hat{\alpha}$	Exact 95% CI	
Empirical	100	0.054	0.0408	0.0699
	1000	0.048	0.0356	0.0631
	5000	0.04	0.0287	0.0541
Ap. Empirical	100	0.047	0.0347	0.062
	1000	0.053	0.0399	0.0688
	5000	0.058	0.0443	0.0743

Table 23. $d = 2$, T_{10}

Type	n	$\hat{\alpha}$	Exact 95% CI	
Empirical	100	0.067	0.0523	0.0843
	1000	0.043	0.0313	0.0575
	5000	0.051	0.0382	0.0665
Ap. Empirical	100	0.067	0.0523	0.0843
	1000	0.035	0.0245	0.0483
	5000	0.044	0.0321	0.0586

The estimated size for the approximate empirical test under $n = 100$ is unacceptably high in the standardized exponential(1) case (recorded in Table 24). But under the following higher sample sizes the approximate empirical test behaves fine. All of the empirical tests have appropriate sizes in the standardized exponential(1) context.

Table 24. $d = 2$, Standardized Exponential(1)

Type	n	$\hat{\alpha}$	Exact 95% CI	
Empirical	100	0.058	0.0443	0.0743
	1000	0.047	0.0347	0.062
	5000	0.052	0.0391	0.0676
Ap. Empirical	100	0.081	0.0648	0.0997
	1000	0.048	0.0356	0.0631
	5000	0.049	0.0365	0.0643

We conclude that the empirical and approximate empirical tests behave well in testing $\mathrm{H}_0 : d = 2$. Deviations from the true size are minor, or remedied by an accurate size in the opposing testing method.

We add covariance differences to the $d = 2$ setting to obtain a $d = 4$ situation. The raised variances in group 3 allow for covariances of larger magnitude. We add these and switch the signs of some of covariances. All entries involve x_8. We actually change the covariance of x_8 with 4 different variables, $x_1, x_3, x_4,$ and x_5.

The group 3 covariance column corresponding to x_8 will differ from the others. This will provide one new difference dimension. It may seem like the differences from the other groups in the covariance row of x_8 in group 3 should each add a separate difference dimension to the SMVCIR space. However, all of these differences in the

covariance row of x_8 are linearly independent. So they only add 1 unique difference dimension to the SMVCIR space.

This is the new covariance matrix for group 3. The entries that differ from the $d = 2$ covariance matrix are marked in boldface.

$$\boldsymbol{\Sigma_{43}} = \begin{bmatrix} 20 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{-16} \\ 0 & 22 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 9 & 7 & -2 & 0 & 0 & \mathbf{-3} \\ 0 & 0 & 7 & 11 & -3 & 0 & 0 & \mathbf{-6} \\ 0 & 0 & -2 & -3 & 10 & 0 & 0 & \mathbf{-1} \\ 0 & 0 & 0 & 0 & 0 & 19 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 30 & 0 \\ \mathbf{-16} & 0 & \mathbf{-3} & \mathbf{-6} & \mathbf{-1} & 0 & 0 & 24 \end{bmatrix} \tag{6.11}$$

The other parameters remain the same.

$$\begin{aligned} \boldsymbol{\mu_{41}} &= \boldsymbol{\mu_{21}} \\ \boldsymbol{\mu_{42}} &= \boldsymbol{\mu_{22}} \\ \boldsymbol{\mu_{43}} &= \boldsymbol{\mu_{23}} \\ \boldsymbol{\Sigma_{41}} &= \boldsymbol{\Sigma_{21}} \\ \boldsymbol{\Sigma_{42}} &= \boldsymbol{\Sigma_{22}} \end{aligned} \tag{6.12}$$

We have the following eigenvalues.

$$\begin{aligned} \lambda_{41} &= 1.2471 \\ \lambda_{42} &= 0.6264 \\ \lambda_{43} &= 0.2862 \\ \lambda_{44} &= 0.2321 \\ \lambda_{45} &= \ldots = \lambda_{48} = 0 \end{aligned} \tag{6.13}$$

For $d = 4$ empirical testing in the normal setting yields accurate test sizes under all sample size values. These results are recorded in Table 25. The approximate empirical test behaves adequately also. Note the high size of .068 under $n = 5000$.

Table 25. $d = 4$, Normal

Type	n	$\hat{\alpha}$	Exact 95% CI	
Empirical	100	0.047	0.0347	0.062
	1000	0.043	0.0313	0.0575
	5000	0.051	0.0382	0.0665
Ap. Empirical	100	0.051	0.0382	0.0665
	1000	0.044	0.0321	0.0586
	5000	0.068	0.0532	0.0854

The size should become closer to .05 as we increase the sample size n. This may be a case in which the approximate empirical is not the best model for the null reference distribution in smaller sample sizes. We note that the confidence interval is not terribly far from including .05. The lower bound is less than .003 above it.

In the T_{10} context we see more mixed results. These are recorded in Table 26. None are awful though. The confidence interval for the approximate empirical at $n = 1000$ misses the correct size .05 by less than .002. For the empirical test at $n = 5000$, the confidence interval misses the correct size .05 by less than .003.

The standardized exponential(1) case yields good size results for both the empirical and approximate empirical tests. These results are recorded in Table 27.

Table 26. $d = 4$, T_{10}

Type	n	$\hat{\alpha}$	Exact 95% CI	
Empirical	100	0.032	0.022	0.0449
	1000	0.047	0.0347	0.062
	5000	0.034	0.0237	0.0472
Ap. Empirical	100	0.056	0.0426	0.0721
	1000	0.035	0.0245	0.0483
	5000	0.063	0.0487	0.0799

Table 27. $d = 4$, Standardized Exponential(1)

Type	n	$\hat{\alpha}$	Exact 95% CI	
Empirical	100	0.05	0.0373	0.0654
	1000	0.054	0.0408	0.0699
	5000	0.047	0.0347	0.062
Ap. Empirical	100	0.059	0.0452	0.0754
	1000	0.047	0.0347	0.062
	5000	0.048	0.0356	0.0631

We find the behavior of the empirical and approximate empirical tests in testing $H_0 : d = 4$ is acceptable. As in the previous case, deficiencies in one test are remedied by the other. The deficiencies we find are not egregious anyway. In the next section we will do a more detailed comparison of the two tests by examining the power of each under alternative hypotheses, not just their size.

The $d = 5$ case is obtained from the $d = 4$ case by changing the variances of group 2. This yields a new group 2 covariance matrix.

$$\boldsymbol{\Sigma_{52}} = \begin{bmatrix} \mathbf{50} & 0 & 0 & 0 & 0 & 0 & 0 & 6 \\ 0 & \mathbf{42} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mathbf{12} & 7 & -2 & 0 & 0 & 3 \\ 0 & 0 & 7 & \mathbf{10} & -3 & 0 & 0 & 6 \\ 0 & 0 & -2 & -3 & \mathbf{9} & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & \mathbf{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{2} & 0 \\ 6 & 0 & 3 & 6 & 1 & 0 & 0 & \mathbf{34} \end{bmatrix} \tag{6.14}$$

The other parameters remain the same.

$$\begin{aligned} \boldsymbol{\mu_{51}} &= \boldsymbol{\mu_{41}} \\ \boldsymbol{\mu_{52}} &= \boldsymbol{\mu_{42}} \\ \boldsymbol{\mu_{53}} &= \boldsymbol{\mu_{43}} \\ \boldsymbol{\Sigma_{51}} &= \boldsymbol{\Sigma_{41}} \\ \boldsymbol{\Sigma_{53}} &= \boldsymbol{\Sigma_{43}} \end{aligned} \tag{6.15}$$

This results in the following eigenvalues of the SMVCIR kernel.

$$\begin{aligned} \lambda_{51} &= 0.9714 \\ \lambda_{52} &= 0.8715 \\ \lambda_{53} &= 0.3848 \\ \lambda_{54} &= 0.1753 \\ \lambda_{55} &= 0.1085 \\ \lambda_{56} &= \lambda_{57} = \lambda_{58} = 0 \end{aligned} \tag{6.16}$$

In the normal population setting, we see excellent size results for both the empirical and approximate empirical tests (recorded in Table 28). The size is low for both when $n = 100$, but it is close to the correct .05 at the two high sample sizes $n = 1000, 5000$. The accuracy of the tests is asymptotic, so the low size at $n = 100$ is not concerning at all.

We see generally good results for the T_{10} setting as well. These results are recorded in Table 29. The test size is too low when $n = 100$. Also in the approximate empirical test at $n = 1000$, the size is somewhat low as well. But otherwise both the empirical and approximate empirical behave well. Most importantly, the size at $n = 5000$ appears accurate.

Table 28. $d = 5$, Normal

Type	n	$\hat{\alpha}$	Exact 95% CI	
Empirical	100	0.021	0.013	0.0319
	1000	0.051	0.0382	0.0665
	5000	0.05	0.0373	0.0654
Ap. Empirical	100	0.025	0.0162	0.0367
	1000	0.057	0.0435	0.0732
	5000	0.055	0.0417	0.071

The standardized exponential(1) setting results are similar to those of the T_{10} case. They are recorded in Table 30. The size for $n = 100$ is small on both tests. Additionally, as in the last setting, the approximate empirical test for $n = 1000$ is a bit too low. Under $n = 5000$, the confidence intervals for both the approximate and approximate empirical include the correct size .05.

Table 29. $d = 5$, T_{10}

Type	n	$\hat{\alpha}$	Exact 95% CI	
Empirical	100	0.031	0.0212	0.0437
	1000	0.04	0.0287	0.0541
	5000	0.047	0.0347	0.062
Ap. Empirical	100	0.025	0.0162	0.0367
	1000	0.034	0.0237	0.0472
	5000	0.057	0.0435	0.0732

Table 30. $d = 5$, Standardized Exponential(1)

Type	n	$\hat{\alpha}$	Exact 95% CI	
Empirical	100	0.016	0.0092	0.0259
	1000	0.051	0.0382	0.0665
	5000	0.06	0.0461	0.0766
Ap. Empirical	100	0.014	0.0077	0.0234
	1000	0.035	0.0245	0.0483
	5000	0.054	0.0408	0.0699

Our final mean, variance, and covariance context has an SMVCIR dimension of $d = 6$. We obtain it from the $d = 5$ setting by changing the mean of group 3.

$$\boldsymbol{\mu}'_{\mathbf{63}} = \left[\begin{array}{cccccccc} \mathbf{-5} & -5 & \mathbf{-5} & 2 & \mathbf{-5} & 6 & 7 & 0 \end{array} \right]' \tag{6.17}$$

The other parameters remain the same.

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{61}} &= \boldsymbol{\mu}_{\mathbf{51}} \\ \boldsymbol{\mu}_{\mathbf{62}} &= \boldsymbol{\mu}_{\mathbf{52}} \\ \boldsymbol{\Sigma}_{\mathbf{61}} &= \boldsymbol{\Sigma}_{\mathbf{51}} \\ \boldsymbol{\Sigma}_{\mathbf{62}} &= \boldsymbol{\Sigma}_{\mathbf{52}} \\ \boldsymbol{\Sigma}_{\mathbf{63}} &= \boldsymbol{\Sigma}_{\mathbf{53}} \end{aligned} \tag{6.18}$$

This results in the following eigenvalues of the SMVCIR kernel.

$$\begin{aligned} \lambda_{61} &= 1.6780 \\ \lambda_{62} &= 0.7451 \\ \lambda_{63} &= 0.4213 \\ \lambda_{64} &= 0.2761 \\ \lambda_{65} &= 0.1455 \\ \lambda_{66} &= 0.0962 \\ \lambda_{67} &= \lambda_{68} = 0 \end{aligned} \tag{6.19}$$

We obtain good test size results for both empirical and the approximate empirical in the normal population setting. They are recorded in Table 31. The size is too small for $n = 100$, but it is close to the correct .05 at the higher sample sizes.

There are also good results for the T_{10} population setting, recorded in Table 32. The size is too small for $n = 100$ again. Also, the size under the approximate empirical is too small for $n = 1000$. But at $n = 5000$ both empirical and approximate empirical are close to the correct size .05.

Table 31. $d = 6$, Normal

Type	n	$\hat{\alpha}$	Exact 95% CI	
Empirical	100	0.027	0.0179	0.039
	1000	0.051	0.0382	0.0665
	5000	0.06	0.0461	0.0766
Ap. Empirical	100	0.03	0.0203	0.0426
	1000	0.051	0.0382	0.0665
	5000	0.052	0.0391	0.0676

Table 32. $d = 6$, T_{10}

Type	n	$\hat{\alpha}$	Exact 95% CI	
Empirical	100	0.034	0.0237	0.0472
	1000	0.055	0.0417	0.071
	5000	0.053	0.0399	0.0688
Ap. Empirical	100	0.031	0.0212	0.0437
	1000	0.033	0.0228	0.046
	5000	0.053	0.0399	0.0688

In the standardized exponential(1) case we see generally good results. They are recorded in Table 33. At $n = 100$ the test sizes are too small. this is also the case for the empirical test at $n = 5000$, but the confidence interval is within .002 of the correct test size .05.

Table 33. $d = 6$, Standardized Exponential(1)

Type	n	$\hat{\alpha}$	Exact 95% CI	
Empirical	100	0.019	0.0115	0.0295
	1000	0.05	0.0373	0.0654
	5000	0.035	0.0245	0.0483
Ap. Empirical	100	0.015	0.0084	0.0246
	1000	0.047	0.0347	0.062
	5000	0.05	0.0373	0.0654

We have evaluated the empirical and approximate empirical tests under the 6 null distribution and 3 fairly different random variable population settings. Our results are quite good. They do not indicate that the empirical and approximately empirical are completely equivalent to each other, but that they are both good tests under the null hypothesis.

We will now see how good they are under alternative hypotheses.

2. Test Power

We first examined the power of the empirical and approximate empirical using the previous $d > d_{null}$ situations as alternatives to the null $d = d_{null}$. At the smallest tested sample size, $n = 100$, we saw some variation in power. But this may have been due to the test statistics non-convergence at this low sample size. In the last

section, we saw that in many cases this sample size showed size that was much lower than the accurate .05. So examination of the power to reject the alternative may be inappropriate at this low sample size.

At higher sample sizes, the alternatives were rejected almost all the time. In comparison to each other, the $d = i$ situations in the last section represent significantly different situations. Our tests detect this fact and indicate that they are truly different. This is encouraging and shows that the test have adequate power when the alternative situation is very different. However, it does not show us how sensitive the tests are, or how they differ from each other.

To determine how the tests behave in the presence of small differences in the alternative from the null, we formulate three new situations.

In the first situation, we examine the power β_1, for the alternative $d = 1$ versus the null $d = 0$. We take our $d = 0$ situation from the last section, and change the mean vector for group 1 very slightly. This creates a new $d = 1$ situation.

This is the new group 1 mean parameter.

$$\boldsymbol{\mu}'_{\mathbf{011}} = \begin{bmatrix} \mathbf{4} & -5 & 3 & 2 & 1 & 6 & \mathbf{6} & 0 \end{bmatrix}' \tag{6.20}$$

We have shown how it differs from the others by marking the differing locations in boldface. We index the new parameters with a null, alternative, group system to differentiate them from the parameters in the last section.

$$\boldsymbol{\mu}'_{\mathbf{012}} = \boldsymbol{\mu}'_{\mathbf{013}} = \begin{bmatrix} 5 & -5 & 3 & 2 & 1 & 6 & 7 & 0 \end{bmatrix}' \tag{6.21}$$

The variance matrices are all the same and identical to their $d = 0$ case.

$$\begin{aligned} \mathbf{\Sigma_{011}} &= \mathbf{\Sigma_{01}} \\ \mathbf{\Sigma_{012}} &= \mathbf{\Sigma_{02}} \\ \mathbf{\Sigma_{013}} &= \mathbf{\Sigma_{03}} \end{aligned} \tag{6.22}$$

These are the eigenvalues for the new situation.

$$\begin{aligned} \lambda_{011} &= 0.03680982 \\ \lambda_{012} &= \ldots = \lambda_{018} = 0 \end{aligned} \tag{6.23}$$

We simulate 1000 independent $n = 1000$ samples from this setting in each of the normal, T_{10}, and standardized exponential(1) population contexts. At this sample size the empirical and approximate empirical both mostly had the correct sizes. The test statistic for $H_0 : d = 0$ is calculated for each sample.

Using our size results from the last chapter and our new test statistic samples, we can evaluate the approximate empirical and empirical tests according to Lloyd's methodology. We discussed this in Chapter V section A. Our B_0 and B_1 values are both 1000. We start by examining the Normal setting results in Table 34.

Table 34. $H_0 : d = 0$ vs. $H_1 : d = 1$, Normal

	Ap. Empirical		Empirical	
	Est.	S.E.	Est.	S.E.
α	0.047	0.0068	0.055	0.0073
β	0.227	0.0141	0.226	0.0141
$\beta - \alpha$	0.18	0.0157	0.171	0.0159
δ	0.9259	0.0834	0.8461	0.0807
$W(0,1)$	0.7437	0.019	0.7252	0.019
$R(.05)$	0.2361	0.0257	0.2122	0.0234

In the normal setting, we find little difference in the power between the two tests. Adjusting for size we see a difference of .01 in $\beta - \alpha$. But when we consider the standard errors for the difference it disappears. This is true for the other differences as well, exempting the actual test size α. So we conclude that when the test power is taken into account, the difference between the empirical and approximate empirical is negligible in this normal population setting. Now we examine the T_{10} setting results in Table 35.

Table 35. $H_0 : d = 0$ vs. $H_1 : d = 1$, T_{10}

	Ap. Empirical		Empirical	
	Est.	S.E.	Est.	S.E.
α	0.033	0.0057	0.038	0.0061
β	0.095	0.0095	0.097	0.0096
$\beta - \alpha$	0.062	0.0111	0.059	0.0114
δ	0.5278	0.0957	0.4755	0.0927
$W(0,1)$	0.6455	0.0252	0.6317	0.0247
$R(.05)$	0.132	0.0205	0.1211	0.0187

We have a similar situation when we move to the T_{10} population setting. Here the power is lower, but the empirical and approximate empirical tests are still similar in power to each other. We can also use the results in Table 34 and Table 35 to contrast the behavior of the empirical (or approximate empirical) under different population settings. Clearly both perform better under the normal situation. The δ parameter represents the intrinsic power of the test. Even factoring in the standard errors of its estimation, it is clearly larger in the normal case. The $R(.05)$ measure, giving an estimate of the power given the correct size is also clearly larger in the normal case. Finally the proportion of correct classifications parameter, $W(0,1)$ is better in the normal case as well. Now we will examine the exponential case in Table 36.

Table 36. $H_0 : d = 0$ vs. $H_1 : d = 1$, Standardized Exponential(1)

	Ap. Empirical		Empirical	
	Est.	S.E.	Est.	S.E.
α	0.057	0.0074	0.049	0.0069
β	0.123	0.0107	0.124	0.0108
$\beta - \alpha$	0.066	0.0131	0.075	0.0128
δ	0.4203	0.0837	0.4994	0.0861
$W(0,1)$	0.6169	0.0226	0.638	0.0228
$R(.05)$	0.1104	0.0158	0.126	0.0178

In the standardized exponential case, we see again see similarity between the empirical and approximate empirical. There are slightly more significant differences between the two, but they are still minor. In comparison to the other population cases, on the objective measures of power δ, $W(0,1)$, and $R(.05)$ it is roughly equivalent to the T_{10} case. It is thus inferior to the normal case.

This indicates that the skewness of the standardized exponential setting and the heavy tails of the students T settings are both detrimental to the empirical and approximate empirical tests. These tests do perform well in these situations, but they perform better in the strict normal case. We also note that this is for detection of a small location difference between groups. Now we will see how the tests perform when detecting a small variance difference.

In this situation, we examine the power β_2, for the alternative $d = 2$ versus the null $d = 1$. We take our $d = 1$ situation from the last section, and change the variance for group 3 very slightly.

This is the new group 3 variance matrix. The changed entries are given in

boldface.

$$\boldsymbol{\Sigma_{123}} = \begin{bmatrix} \mathbf{13} & 0 & 0 & 0 & 0 & 0 & 0 & 6 \\ 0 & 12 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 9 & 7 & -2 & 0 & 0 & 3 \\ 0 & 0 & 7 & 11 & -3 & 0 & 0 & 6 \\ 0 & 0 & -2 & -3 & 10 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 9 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{15} & 0 \\ 6 & 0 & 3 & 6 & 1 & 0 & 0 & \mathbf{16} \end{bmatrix} \tag{6.24}$$

$$\begin{aligned} \boldsymbol{\mu'_{121}} &= \boldsymbol{\mu'_{11}} \\ \boldsymbol{\mu'_{122}} &= \boldsymbol{\mu'_{12}} \\ \boldsymbol{\mu'_{123}} &= \boldsymbol{\mu'_{13}} \end{aligned} \tag{6.25}$$

$$\boldsymbol{\Sigma_{121}} = \boldsymbol{\Sigma_{122}} = \begin{bmatrix} 10 & 0 & 0 & 0 & 0 & 0 & 0 & 6 \\ 0 & 12 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 9 & 7 & -2 & 0 & 0 & 3 \\ 0 & 0 & 7 & 11 & -3 & 0 & 0 & 6 \\ 0 & 0 & -2 & -3 & 10 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 9 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 10 & 0 \\ 6 & 0 & 3 & 6 & 1 & 0 & 0 & 14 \end{bmatrix} \tag{6.26}$$

This results in the following eigenvalues of the SMVCIR kernel.

$$\begin{aligned} \lambda_{121} &= 1.165091 \\ \lambda_{122} &= 0.01078596 \\ \lambda_{123} &= \lambda_{128} = 0 \end{aligned} \tag{6.27}$$

As before, we simulate 1000 $n = 1000$ samples from this mean, variance, covariance context. We use the same 3 population settings as before. The power estimates are taken from the new samples, and the size estimates are taken from the last section in the case $d = 1$.

We start by examining the Normal setting results in Table 37. We see little dif-

Table 37. $H_0 : d = 1$ vs. $H_1 : d = 2$, Normal

	Ap. Empirical		Empirical	
	Est.	S.E.	Est.	S.E.
α	0.047	0.0068	0.049	0.0069
β	0.085	0.009	0.106	0.01
$\beta - \alpha$	0.038	0.0113	0.057	0.0122
δ	0.3025	0.0901	0.4065	0.0873
$W(0,1)$	0.5847	0.0248	0.6131	0.0236
$R(.05)$	0.0897	0.0146	0.1078	0.0162

ference between the empirical and approximate empirical tests in the normal setting. Both tests perform fairly well.

Both tests perform well in the T_{10} setting as well. This is shown in Table 38 The powers of both tests are similar again, and we see that the objective measures are within a standard deviation of each other across the two tests. The tests in this setting are closer to the tests under the normal setting than in our first power setting (where we test $d = 0$ vs. $d = 1$). They appear to be roughly equivalent in quality in this new mean, variance, covariance test setting.

We see the same results under the standardized exponential setting in Table 39. This tells us that the approximate empirical and and empirical tests behave well in testing a small variance difference across a fairly broad range of distributions.

Table 38. $H_0 : d = 1$ vs. $H_1 : d = 2$, T_{10}

	Ap. Empirical		Empirical	
	Est.	S.E.	Est.	S.E.
α	0.032	0.0056	0.046	0.0067
β	0.069	0.0082	0.069	0.0082
$\beta - \alpha$	0.037	0.0099	0.023	0.0106
δ	0.3689	0.0994	0.2017	0.0928
$W(0,1)$	0.6029	0.0271	0.5567	0.0259
$R(.05)$	0.101	0.0176	0.0745	0.0131

Table 39. $H_0 : d = 1$ vs. $H_1 : d = 2$, Standardized Exponential(1)

	Ap. Empirical		Empirical	
	Est.	S.E.	Est.	S.E.
α	0.043	0.0065	0.048	0.0068
β	0.093	0.0094	0.073	0.0084
$\beta - \alpha$	0.05	0.0114	0.025	0.0108
δ	0.3944	0.0908	0.2108	0.0914
$W(0,1)$	0.6098	0.0246	0.5592	0.0255
$R(.05)$	0.1056	0.0166	0.0758	0.013

Now we will examine how the empirical and approximate empirical tests behave in detecting a small covariance difference. We begin with the $d = 2$ setting of the last section, and slightly change the covariance in group 3 in for 2 separate variable duos.

$$\boldsymbol{\Sigma_{243}} = \begin{bmatrix} 20 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{3} \\ 0 & 22 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 9 & 7 & -2 & 0 & 0 & 3 \\ 0 & 0 & 7 & 11 & -3 & 0 & 0 & 6 \\ 0 & 0 & -2 & -3 & 10 & 0 & 0 & \mathbf{0} \\ 0 & 0 & 0 & 0 & 0 & 19 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 30 & 0 \\ \mathbf{3} & 0 & 3 & 6 & \mathbf{0} & 0 & 0 & 24 \end{bmatrix} \tag{6.28}$$

These are the other parameters.

$$\boldsymbol{\Sigma_{241}} = \boldsymbol{\Sigma_{242}} = \begin{bmatrix} 10 & 0 & 0 & 0 & 0 & 0 & 0 & 6 \\ 0 & 12 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 9 & 7 & -2 & 0 & 0 & 3 \\ 0 & 0 & 7 & 11 & -3 & 0 & 0 & 6 \\ 0 & 0 & -2 & -3 & 10 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 9 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 10 & 0 \\ 6 & 0 & 3 & 6 & 1 & 0 & 0 & 14 \end{bmatrix} \tag{6.29}$$

$$\begin{aligned} \boldsymbol{\mu_{241}} &= \boldsymbol{\mu_{21}} \\ \boldsymbol{\mu_{243}} = \boldsymbol{\mu_{242}} &= \boldsymbol{\mu_{22}} \end{aligned} \tag{6.30}$$

The eigenvalues of the new setting are the following.

$$\begin{aligned} \lambda_{241} &= 1.019897 \\ \lambda_{242} &= 0.3458705 \\ \lambda_{243} &= 0.0005687747 \\ \lambda_{244} &= 0.0004613893 \\ \lambda_{245} &= \ldots = \lambda_{248} = 0 \end{aligned} \tag{6.31}$$

We begin with the Normal population setting in Table 40.

Table 40. $H_0 : d = 2$ vs. $H_1 : d = 4$, Normal

	Ap. Empirical		Empirical	
	Est.	S.E.	Est.	S.E.
α	0.053	0.0072	0.048	0.0068
β	0.106	0.01	0.137	0.0113
$\beta - \alpha$	0.053	0.0123	0.089	0.0132
δ	0.3684	0.0861	0.5707	0.0857
$W(0,1)$	0.6027	0.0235	0.6567	0.0223
$R(.05)$	0.1009	0.0152	0.1414	0.0192

Both the empirical and approximate empirical give similar results for the normal setting. We note that the empirical test appears to be slightly more powerful. This is not true for just the raw power β, but on the more sophisticated measures such as the coefficient of determination $W(0,1)$.

The empirical and approximate empirical are roughly equivalent under the T_{10} population setting. These results are recorded in Table 41. Both are slightly less powerful than under the normal case. The standardized exponential setting (Table 42) does not differ very much from the other two population settings. The empirical and approximate empirical are roughly equivalent here also.

Table 41. $H_0 : d = 2$ vs. $H_1 : d = 4$, T_{10}

	Ap. Empirical		Empirical	
	Est.	S.E.	Est.	S.E.
α	0.035	0.0059	0.043	0.0065
β	0.076	0.0085	0.086	0.0091
$\beta - \alpha$	0.041	0.0104	0.043	0.0111
δ	0.3794	0.0966	0.3511	0.0915
$W(0,1)$	0.6058	0.0263	0.598	0.025
$R(.05)$	0.1029	0.0173	0.0979	0.0158

Table 42. $H_0 : d = 2$ vs. $H_1 : d = 4$, Standardized Exponential(1)

	Ap. Empirical		Empirical	
	Est.	S.E.	Est.	S.E.
α	0.048	0.0068	0.047	0.0068
β	0.084	0.009	0.082	0.0089
$\beta - \alpha$	0.036	0.0113	0.035	0.0112
δ	0.2859	0.0899	0.2829	0.0905
$W(0,1)$	0.5801	0.0248	0.5793	0.025
$R(.05)$	0.0871	0.0142	0.0866	0.0143

Together, these results suggest that the empirical and approximate empirical may exhibit only negligible differences in the presence of detecting a very weak alternative. We do see that the distributional setting can matter in the power of the tests, as we saw in the $d = 0$ vs. $d = 1$ case. And the slight but noticeable difference we saw between empirical and approximate empirical tests in the $d = 2$ vs. $d = 4$ case under the normal population may hint at a true significant difference between the two tests in certain situations. Our results are not exhaustive, but we have shown the success of both the empirical and approximate empirical tests in a broad variety of contexts.

We have not yet shown how the empirical and the approximate empirical tests behave in the iterated testing scheme (4.1) that we use to choose our SMVCIR dimension. We do this in the next section.

3. Choose d SMVCIR, SIR, and SAVE

We simulated 1000 $n = 5000$ samples from each of the $d = 0, 1, 2, 4, 5, 6$ mean, variance, covariance settings applied to each of our three population settings. We defined each of these parameter settings in the first section. The population settings were defined in the chapter's introduction.

Using a size .05 decision rule for each test performed, we implemented the iterated testing scheme (4.1) for each of SMVCIR, SAVE, and SIR. As we found in the first section, the test size of the individual tests should be close to .05 due to the $n = 5000$ sample size. But as we mentioned earlier, these are informal simulations. We will not be greatly concerned over the individual test size. Our criterion for success is that the proper dimension is chosen reasonably often.

To ensure independence of the empirical and approximate empirical tests in our results, we performed separate simulations for each. SAVE and SIR were computed in other separate simulations. We start our study in Table 43 with $d = 0$ imposed on

the Normal population setting.

Table 43. d Choice, $d = 0$, Normal

	Ap. Empirical	Empirical	SIR	SAVE
0	944	952	960	959
1	42	39	40	40
2	14	8		1
3		1		
4				
5				
6				
7				
8				

In the normal setting and $d = 0$ case, we see that each algorithm picked the correct dimension approximately 95% of the time. This is encouraging. We find good results for the remainder of the $d = 0$ cases and each of the $d = 1$ cases (as shown in later tables). So each algorithm performed well at determining that there was no group differences ($d = 0$) and that there was a single difference when the difference was in location. The T_{10} setting results for $d = 0$ are recorded in Table 44. The exponential results are recorded in Table 45.

Table 44. d Choice, $d = 0$, T_{10}

	Ap. Empirical	Empirical	SIR	SAVE
0	946	959	939	953
1	48	32	60	46
2	4	8	1	1
3	1	1		
4	1			
5				
6				
7				
8				

Table 45. d Choice, $d = 0$, Standardized Exponential(1)

	Ap. Empirical	Empirical	SIR	SAVE
0	951	944	950	948
1	39	50	48	51
2	8	6	2	1
3	1			
4	1			
5				
6				
7				
8				

The next three tables record how the tests performed under $d = 1$ in our three population settings. The Normal test results are recorded in Table 46. The T_{10} results are recorded in Table 47. Finally the standardized exponential results are contained in Table 48.

Table 46. d Choice, $d = 1$, Normal

	Ap. Empirical	Empirical	SIR	SAVE
0				
1	940	942	946	951
2	46	43	54	48
3	14	14		1
4		1		
5				
6				
7				
8				

Table 47. d Choice, $d = 1$, T_{10}

	Ap. Empirical	Empirical	SIR	SAVE
0				
1	937	945	946	956
2	57	45	54	44
3	4	9		
4	1	1		
5	1			
6				
7				
8				

Table 48. d Choice, $d = 1$, Standardized Exponential(1)

	Ap. Empirical	Empirical	SIR	SAVE
0				
1	955	965	954	962
2	36	35	46	38
3	6			
4	3			
5				
6				
7				
8				

Under the $d = 2$ situation, there is a variance group difference in five of the variables. There is also a group mean difference. SIR should detect only one difference, that of the mean. SIR correctly detects this one difference in each of the three population settings. SMVCIR should detect 2 dimensions. It does this well in all of the population settings and for both the empirical and approximate empirical tests. We see the Normal population results in Table 49.

Table 49. d Choice, $d = 2$, Normal

	Ap. Empirical	Empirical	SIR	SAVE
0				
1			960	
2	938	956	40	
3	48	35		
4	13	8		
5	1	1		237
6				718
7				45
8				

The mean of group 1 differs from the others for predictors x_1, x_2, x_3 and x_7. The variance of predictors x_1, x_2, x_6, x_7, and x_8 in group 3 is different than the other group variances. So we should find 6 difference dimensions under SAVE. Recall that the variance are not stacked in SAVE, so each may contribute a difference.

This difference dimension is chosen most often, but the next highest dimension, five differences is also chosen with some regularity. We calculated the SAVE kernel in this case, and the lowest non-zero eigenvalue (approximately .0168) was one quarter the size of the the next highest eigenvalue at position 5 (approximately .07). The eigenvalue at position 5 was small compared to the other values as well. This happens because the variance differences can be added together to closely approximate the lone mean difference. This small eigenvalue means that the last dimension has low discriminating power, and that the power to detect it may be low.

All our variance differences are practically significant, so this is happening because the variance differences can be added together to closely approximate the lone mean difference.

We will demonstrate how this works using a small 4 predictor example.

Suppose we have the following mean difference vector for group j.

$$\boldsymbol{\mu_j} = \begin{bmatrix} \mu_{j1} \\ \mu_{j2} \\ \mu_{j3} \\ \mu_{j4} \end{bmatrix} \tag{6.32}$$

The variance of group m is different from the other groups for the first 3 predictors. This yields the following variance difference vectors in the $SAVE$ spanning set.

$$\begin{aligned} \boldsymbol{v_{j1}}' &= \begin{bmatrix} v_{j1} & 0 & 0 & 0 \end{bmatrix} \\ \boldsymbol{v_{j2}}' &= \begin{bmatrix} 0 & v_{j2} & 0 & 0 \end{bmatrix} \\ \boldsymbol{v_{j3}}' &= \begin{bmatrix} 0 & 0 & v_{j3} & 0 \end{bmatrix} \end{aligned} \tag{6.33}$$

By using the right coefficients (dividing the by non-zero variance difference and multiplying by the mean difference), we can create a linear form of the variance difference vectors that nearly equals the mean difference vector. The last position μ_{j4}

cannot be approximated by the variance differences, but if it is negligible relative to the other components the approximation will be successful.

$$\boldsymbol{\mu_j} \approx \frac{\mu_{j1}}{v_{j1}}\boldsymbol{v_{j1}} + \frac{\mu_{j2}}{v_{j2}}\boldsymbol{v_{j2}} + \frac{\mu_{j1}}{v_{j3}}\boldsymbol{v_{j3}} \tag{6.34}$$

This process is happening here, and it appears to be successful a significant portion of the time. The same type of situation can occur in SMVCIR when we deal with covariance differences, where a significant number of elements in the difference vectors may be zero. Note that this approximation does not mean that the mean difference is excluded from the spanning set when the approximation works. Under successful approximation, we may have

$$\boldsymbol{\mu_j} - \frac{\mu_{j1}}{v_{j1}}\boldsymbol{v_{j1}} - \frac{\mu_{j2}}{v_{j2}}\boldsymbol{v_{j2}} \approx \frac{\mu_{j1}}{v_{j3}}\boldsymbol{v_{j3}} \tag{6.35}$$

In spot checking of several instances, where the correct dimension of 6 was chosen and also where the lower and incorrect choice of 5 was selected, we found that the mean difference was always selected and placed first in the SAVE dimension ordering. A variance difference was the omitted dimension in each of these cases.

For the T_{10} population setting (Table 50), SMVCIR and SIR behave well. SAVE is nearly equally likely to pick 5 dimensions as 6 dimensions here. As in the normal setting, SAVE's power in detecting the last dimension is low. But why is it lower than the normal. Spot checking of several cases suggested that the final SAVE difference dimension is so small that outliers are masking its detection. We found that several severe outliers in a group other than 2 (2 is the group that differs in variance from the others in this setting) could subtly bloat the variance of the group they belong to and make it equivalent to that of group 2 in this final SAVE dimension. As in the normal case, the mean difference is placed at the beginning of the dimension ordering, so all discrimination for the other dimensions must be in variance.

Table 50. d Choice, $d = 2$, T_{10}

	Ap. Empirical	Empirical	SIR	SAVE
0				
1			968	
2	955	960	32	
3	36	31		
4	6	9		
5	3			498
6				471
7				30
8				1

The T_{10} has heavier tails than the normal distribution, so it will provide more outliers when sampled. Hence we find that the last non-null dimension 6 is detected less often in this setting.

For the standardized exponential(1) setting (Table 51), we see an even greater concentration in the difference dimension 5 choice. We spot checked samples from this setting as well. Again, outliers were masking the discrimination of dimension 6. This distribution yields even more outliers than the T_{10}, so we expect to see even lower power to detect the correct SAVE dimension 6. We find this to be the case.

Table 51. d Choice, $d = 2$, Standardized Exponential(1)

	Ap. Empirical	Empirical	SIR	SAVE
0				
1			967	
2	953	943	33	
3	38	48		
4	6	7		
5	2	2		698
6	1			279
7				21
8				2

Do note that under all the population settings, SAVE's test of the correct null dimension has adequate size. Division of the number of correct choices for $d = 6$ under the normal distribution (718) by itself plus the number of incorrect choices of higher dimensions (718+45), results in a selection rate of 0.9410223. This is quite close to .95, and suggests the the size is close to the correct value of .05. Similarly, we see a proper selection rate of 0.938247 under the T_{10} and 0.923841 under the standardized exponential(1).

This provides further evidence that we should analyze our iterated test algorithm (4.1) as a multiple hypothesis test situation. The performance of SAVE under these population settings is hampered not by the incorrect size and power of the test at the correct and higher incorrect dimensions, but the weak power of the tests for lower incorrect dimensions.

When $d = 4$, covariance differences are added. So we have a mean difference, variance difference, and two covariance differences. In all cases both SIR and SMV-CIR perform as expected. SAVE places these covariance differences together with the variance differences, as they are in the same group. The correct SAVE dimension is 7. Calculation of the SAVE kernel reveals that the final eigenvalue is an order of magnitude below the next lowest. This could be due to the linear approximation problem we discussed previously. It may also be due to low magnitude of the covariance differences. Regardless, as before, the outlier heavy distributions have more difficulty properly detecting $d = 7$. The normal distribution has trouble detecting $d = 7$ as well, but not as severe.

Table 52. d Choice, $d = 4$, Normal

	Ap. Empirical	Empirical	SIR	SAVE
0				
1			962	
2			38	
3				
4	941	950		
5	46	38		
6	12	12		427
7	1			547
8				26

Table 53. d Choice, $d = 4$, T_{10}

	Ap. Empirical	Empirical	SIR	SAVE
0				
1			963	
2			37	
3				
4	945	952		
5	43	34		
6	11	14		647
7	1			331
8				22

We also see appropriate size of the null test for each of the population settings. The normal (Table 52) has a proper selection rate of 0.9546248. The T_{10} (Table 53) had a proper selection rate of 0.937677. The standardized exponential(1) (Table 54) setting had a rate of 0.9392523.

Table 54. d Choice, $d = 4$, Standardized Exponential(1)

	Ap. Empirical	Empirical	SIR	SAVE
0				
1			973	
2			27	
3				
4	946	964		
5	46	29		
6	6	6		786
7	2	1		201
8				13

In the $d = 5$ case, we add a new variance difference. SIR continues to pick a single mean dimension. Both the empirical and approximate empirical tests correctly pick $d = 5$ most of the time. Calculation of the SAVE kernel reveals that it should detect 8 dimensions here. It also shows that the next to last dimension's eigenvalue (approximately .07) is nearly 40 times the eigenvalue of the last dimension (.0018). We find that the power of SAVE to reject the last null hypothesis of (4.1) $H_0 : d = 7$ is low for each population setting. It is especially low under the outlier heavy dimensions.

Table 55. d Choice, $d = 5$, Normal

	Ap. Empirical	Empirical	SIR	SAVE
0				
1			913	
2			87	
3				
4				
5	956	960		
6	35	32		
7	9	8		551
8				449

Table 56. d Choice, $d = 5$, T_{10}

	Ap. Empirical	Empirical	SIR	SAVE
0				
1			916	
2			84	
3				
4				
5	949	951		
6	41	37		
7	10	9		687
8		3		313

Note that SAVE is restricted to mainly 2 dimension choices in the normal (Table 55) and T_{10} (Table 56) settings. But the test power is weak enough and presence of outliers heavy enough in the exponential setting (Table 57) that up to 4 dimensions may be chosen.

Table 57. d Choice, $d = 5$, Standardized Exponential(1)

	Ap. Empirical	Empirical	SIR	SAVE
0				
1			916	
2			84	
3				
4				
5	950	944		3
6	35	47		29
7	14	9		832
8	1			136

To obtain the $d = 6$ case, we add another mean difference. This finally causes SIR to detect 2 differences. SMVCIR correctly detects the right number of dimensions under both the empirical and approximate empirical methods. SAVE has problems of power again.

Table 58. d Choice, $d = 6$, Normal

	Ap. Empirical	Empirical	SIR	SAVE
0				
1				
2			1000	
3				
4				
5				
6	955	946		3
7	37	50		800
8	8	4		197

Table 59. d Choice, $d = 6$, T_{10}

	Ap. Empirical	Empirical	SIR	SAVE
0				
1				
2			1000	
3				
4				
5				
6	957	947		41
7	39	46		820
8	4	7		139

Though the differences in detection between the populations are less pronounced, we still see higher power in the normal (Table 58) than the outlier heavy distributions, T_{10} (Table 59) and standardized exponential (Table 60).

Table 60. d Choice, $d = 6$, Standardized Exponential(1)

	Ap. Empirical	Empirical	SIR	SAVE
0				
1				
2			1000	
3				
4				
5				
6	954	961		193
7	34	34		692
8	12	5		115

These results show that SIR and SMVCIR can effectively pick the correct number of dimension in practice. They also show how SAVE can have difficulty choosing a single dimension when the differences it finds are not strongly linearly independent or of sufficient individual strength. They also demonstrate how outlier heavy distributions can mask slight group differences that light outlier distributions (normal) show with clarity.

These results also draw our attention to the need to analyze these algorithms in a multiple testing setting. Our SMVCIR results only showed type 1 errors an appropriate 5 percent of the time. But some of the discrepancies we encountered while using SAVE may be partially fixed by a more accurate choice of test size.

Using a .05 size for each test may lead to the sequence of tests having a size that is significantly larger. We will discuss this again in the last chapter as we look at further directions of research.

B. Theory

To analyze accuracy the theory (4.27) reference distribution, we will use normal and mixture normal population data. We make draws from this data and compute our test statistic (4.2). Then we calculate the reference distribution's parameters and make an identical number of draws from the reference distribution. If the two coincide (based on Q-Q Plot and Kernel Density Plot) we conclude that the theory reference distribution is the distribution of test statistic (4.2) for the given population setting.

As in the last section, our simulation results will not be authoritative. They will only corroborate the accuracy of our theoretical derivations. They will demonstrate the accuracy of our derivations in a fairly wide variety of contexts though.

Our choices of mixture multivariate normal models are taken from Marron and Wand (1992). Specifically, we use their bimodal univariate density as a starting point for our mixture models. We create new mixtures by altering this model.

We will create mixture normal population settings and invoke SMVCIR on these settings. The marginal (marginalizing over subgroup) group means and variances will then be used as parameters for non-mixture multivariate normal population settings. We will invoke SMVCIR on these settings as well. The rate of convergence of the SMVCIR test statistic to the theory (4.27) distribution will be compared for the normal and mixture normal cases.

1. Mixture Multivariate Normal

In all of our mixture multivariate normal situations, we have 3 groups and 8 predictors. Each group has two subgroups. These are the group proportions.

$$\begin{aligned} p_1 &= .3 \\ p_2 &= .2 \\ p_3 &= .5 \end{aligned} \tag{6.36}$$

Each of the subgroups has the same population proportion of .5.

$$\begin{aligned} p_{11} &= p_{12} = .5 \\ p_{21} &= p_{22} = .5 \\ p_{31} &= p_{32} = .5 \end{aligned} \tag{6.37}$$

In this first setting, $d = 1$, all subgroup variance matrices are the same.

$$\boldsymbol{\Sigma_{111}} = \ldots = \boldsymbol{\Sigma_{132}} = \\ +0.\bar{1} \begin{bmatrix} +4.0 & +0.\bar{1} & 0 & 0 & 0 & 0 & 0 & 0 \\ +0.\bar{1} & +4.0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & +4.0 & 0 & 0 & 0 & 0 & -2.0 \\ 0 & 0 & 0 & +4.0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & +4.0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & +4.0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & +4.0 & 0 \\ 0 & 0 & -2.0 & 0 & 0 & 0 & 0 & +4.0 \end{bmatrix} \tag{6.38}$$

The means of the first subgroup of both group 1 and group 2 are identical as well.

$$\boldsymbol{\mu'_{111}} = \boldsymbol{\mu'_{121}} = \begin{bmatrix} -1 & 0 & 1 & 2 & 4 & 1.5 & 0 & -1 \end{bmatrix}' \tag{6.39}$$

The same is true for the second subgroup of the first two groups. But this mean

is different from the last, providing the bimodality of the group populations.

$$\boldsymbol{\mu}'_{\mathbf{112}} = \boldsymbol{\mu}'_{\mathbf{122}} = \begin{bmatrix} 1 & 0 & -1 & 4 & 2 & -.5 & -2 & 1 \end{bmatrix}' \tag{6.40}$$

Providing our $d = 1$ condition, there is a group mean difference for the third group from the subgroups of the other two.

$$\boldsymbol{\mu}'_{\mathbf{131}} = \begin{bmatrix} 0 & 0 & 1.5 & 0 & 4 & 1 & -1 & 0 \end{bmatrix}' \tag{6.41}$$

$$\boldsymbol{\mu}'_{\mathbf{132}} = \begin{bmatrix} 2 & 0 & -.5 & 2 & 2 & -1 & -3 & 2 \end{bmatrix}' \tag{6.42}$$

Marginalizing over these subgroup parameters using the techniques discussed in the last chapter, we obtain the following marginal group parameters.

$$\boldsymbol{\mu}'_{\mathbf{11}} = \boldsymbol{\mu}'_{\mathbf{12}} = \begin{bmatrix} 0 & 0 & 0 & 3 & 3 & .5 & -1 & 0 \end{bmatrix}' \tag{6.43}$$

$$\boldsymbol{\mu}'_{\mathbf{13}} = \begin{bmatrix} 1 & 0 & .5 & 1 & 3 & 0 & -2 & 1 \end{bmatrix}' \tag{6.44}$$

$$\boldsymbol{\Sigma}_{\mathbf{11}} = \boldsymbol{\Sigma}_{\mathbf{12}} = \boldsymbol{\Sigma}_{\mathbf{13}} = \begin{bmatrix} +1.\bar{4} & +0.\bar{3} & -1.0 & +1.0 & -1.0 & -1.0 & -1.0 & +1.0 \\ +0.\bar{3} & +0.\bar{4} & 0 & 0 & 0 & 0 & 0 & 0 \\ -1.0 & 0 & +1.\bar{4} & -1.0 & +1.0 & +1.0 & +1.0 & -1.\bar{2} \\ +1.0 & 0 & -1.0 & +1.\bar{4} & -1.0 & -1.0 & -1.0 & +1.0 \\ -1.0 & 0 & +1.0 & -1.0 & +1.\bar{4} & +1.0 & +1.0 & -1.0 \\ -1.0 & 0 & +1.0 & -1.0 & +1.0 & +1.\bar{4} & +1.0 & -1.0 \\ -1.0 & 0 & +1.0 & -1.0 & +1.0 & +1.0 & +1.\bar{4} & -1.0 \\ +1.0 & 0 & -1.\bar{2} & +1.0 & -1.0 & -1.0 & -1.0 & +1.\bar{4} \end{bmatrix} \tag{6.45}$$

The eigen values of the resulting SMVCIR kernel are the following.

$$\begin{aligned} \lambda_{11} &= .9346632 \\ \lambda_{12} &= \ldots = \lambda_{18} = 0 \end{aligned} \tag{6.46}$$

Using R, we performed the simulations described earlier for sample size $n = 100, 1000, 5000$ with 1000 draws from the theory (4.27) and 1000 draws from the actual test statistic distribution. The red line represents the theory (4.27) distribution, and the black represents the actual distribution. In the higher sample sizes beyond $n = 100$, the test statistic and the theory distribution matched quite well.

We begin with a Q-Q plot for $n = 100$ in Figure 20. This is followed by the kernel densities for $n = 100$ in Figure 21 and then the Q-Q plot for $n = 1000$ in Figure 22.

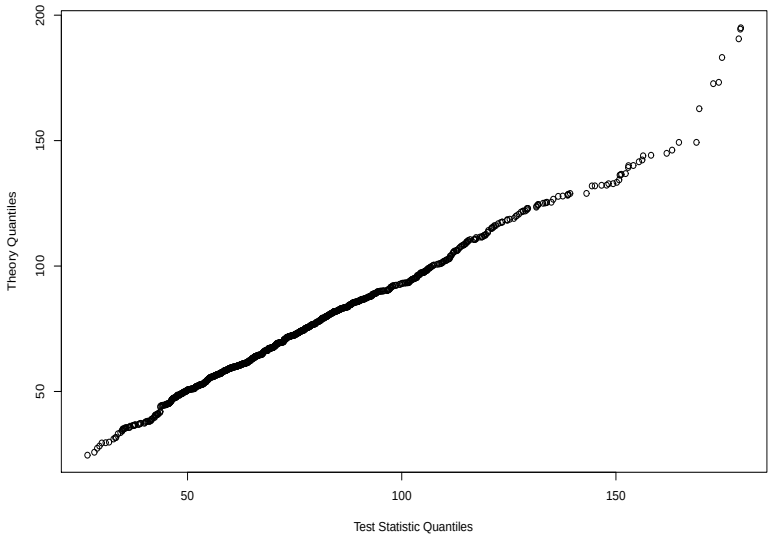

Figure 20. $d = 1$, Mixture Normal, $n = 100$, Q-Q Plot

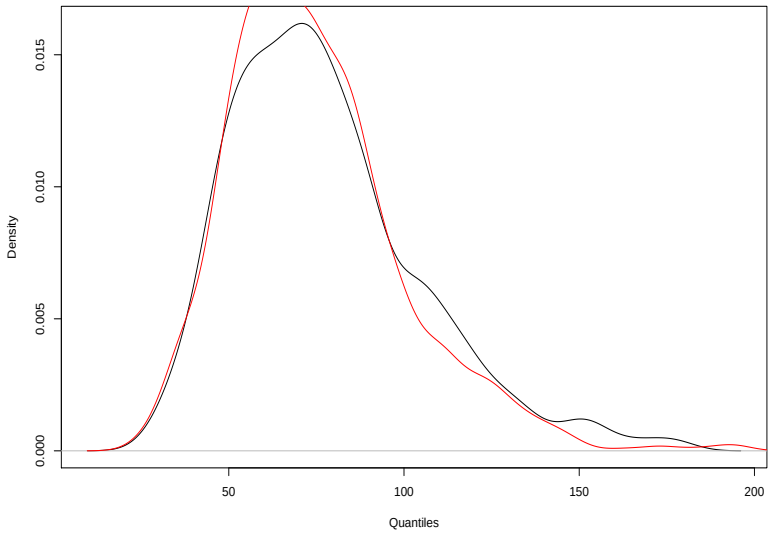

Figure 21. $d = 1$, Mixture Normal, $n = 100$, Kernel Densities

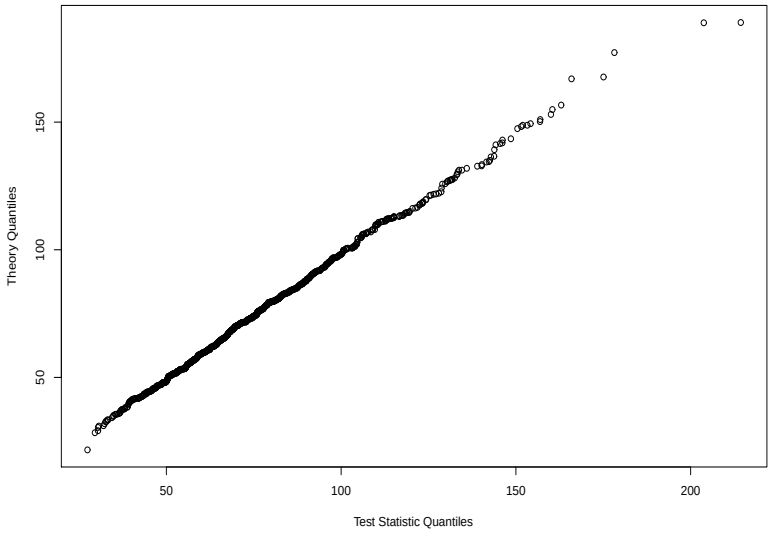

Figure 22. $d = 1$, Mixture Normal, $n = 1000$, Q-Q Plot

Now we examine the kernel densities for $n = 1000$ in Figure 23. We follow with the Q-Q Plot for $n = 5000$ in Figure 24 and the kernel densities for $n = 5000$ in Figure 25.

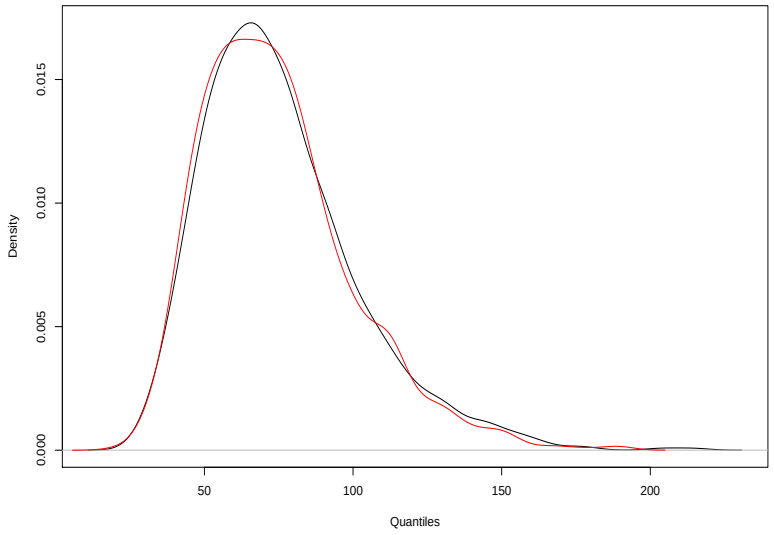

Figure 23. $d = 1$, Mixture Normal, $n = 1000$, Kernel Densities

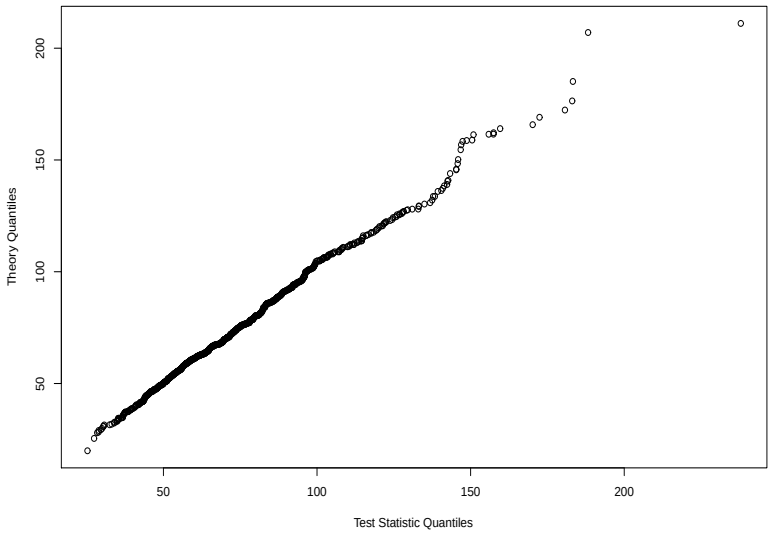

Figure 24. $d = 1$, Mixture Normal, $n = 5000$, Q-Q Plot

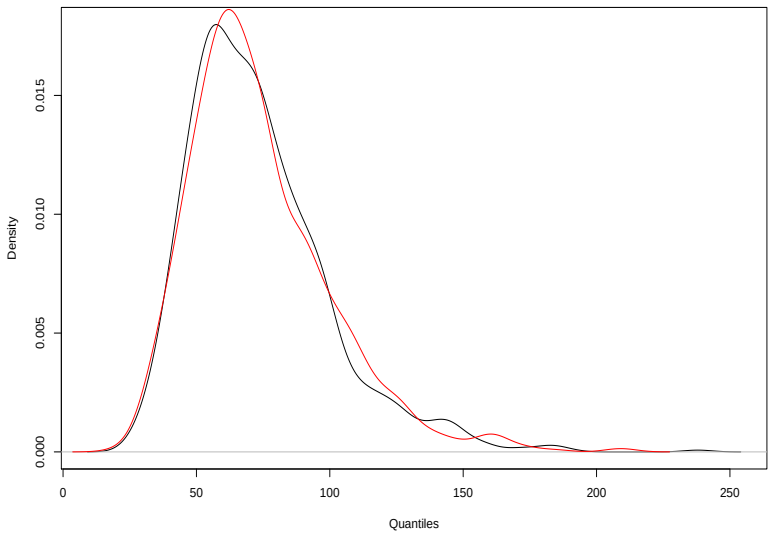

Figure 25. $d = 1$, Mixture Normal, $n = 5000$, Kernel Densities

In the next situation, $d = 2$, as we change the variance of the first subgroup in group 2. The new variance matrix for subgroup 1 group 2 is given below.

$$\boldsymbol{\Sigma_{221}} = \boldsymbol{\Sigma_{121}} - \text{diag}\left(\boldsymbol{\Sigma_{121}}\right) + \begin{bmatrix} 4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 4 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 \end{bmatrix} \tag{6.47}$$

Note that only the variances have been changed the covariances remain the same. This change to the subgroup leads to the following new marginal group variance.

$$\boldsymbol{\Sigma_{22}} = \boldsymbol{\Sigma_{12}} - \text{diag}\left(\boldsymbol{\Sigma_{12}}\right) + \begin{bmatrix} 3.\bar{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 2.\bar{2} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 3.\bar{2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3.\bar{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3.\bar{2} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 3.\bar{2} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 3.\bar{2} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3.\bar{2} \end{bmatrix} \tag{6.48}$$

The other parameters remain the same.

$$
\begin{aligned}
\boldsymbol{\mu_{211}} &= \boldsymbol{\mu_{111}} \\
\boldsymbol{\mu_{212}} &= \boldsymbol{\mu_{112}} \\
\boldsymbol{\mu_{221}} &= \boldsymbol{\mu_{121}} \\
\boldsymbol{\mu_{222}} &= \boldsymbol{\mu_{122}} \\
\boldsymbol{\mu_{231}} &= \boldsymbol{\mu_{131}} \\
\boldsymbol{\mu_{232}} &= \boldsymbol{\mu_{132}} \\
\boldsymbol{\Sigma_{211}} &= \boldsymbol{\mu_{111}} \\
\boldsymbol{\Sigma_{212}} &= \boldsymbol{\mu_{112}} \\
\boldsymbol{\Sigma_{222}} &= \boldsymbol{\mu_{122}} \\
\boldsymbol{\Sigma_{231}} &= \boldsymbol{\mu_{131}} \\
\boldsymbol{\Sigma_{232}} &= \boldsymbol{\mu_{132}} \\
\boldsymbol{\mu_{21}} &= \boldsymbol{\mu_{11}} \\
\boldsymbol{\mu_{22}} &= \boldsymbol{\mu_{12}} \\
\boldsymbol{\mu_{23}} &= \boldsymbol{\mu_{13}} \\
\boldsymbol{\Sigma_{21}} &= \boldsymbol{\Sigma_{11}} \\
\boldsymbol{\Sigma_{23}} &= \boldsymbol{\Sigma_{13}}
\end{aligned}
\tag{6.49}
$$

This situation yields the following kernel eigenvalues.

$$
\begin{aligned}
\lambda_{21} &= 1.664305 \\
\lambda_{22} &= 0.7890368 \\
\lambda_{23} &= \ldots = \lambda_{28} = 0
\end{aligned}
\tag{6.50}
$$

We performed similar simulations to those made for the $d = 1$. True convergence appeared to happen by $n = 5000$. Reasonably good results were found for the lower sample sizes. We begin by examining the Q-Q Plot for $n = 100$ in Figure 26. The kernel densities for $n = 100$ are given in Figure 27.

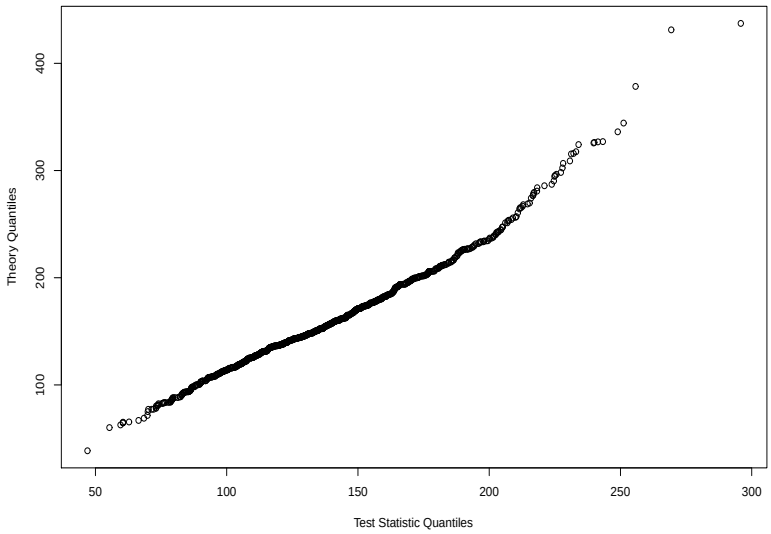

Figure 26. $d = 2$, Mixture Normal, $n = 100$, Q-Q Plot

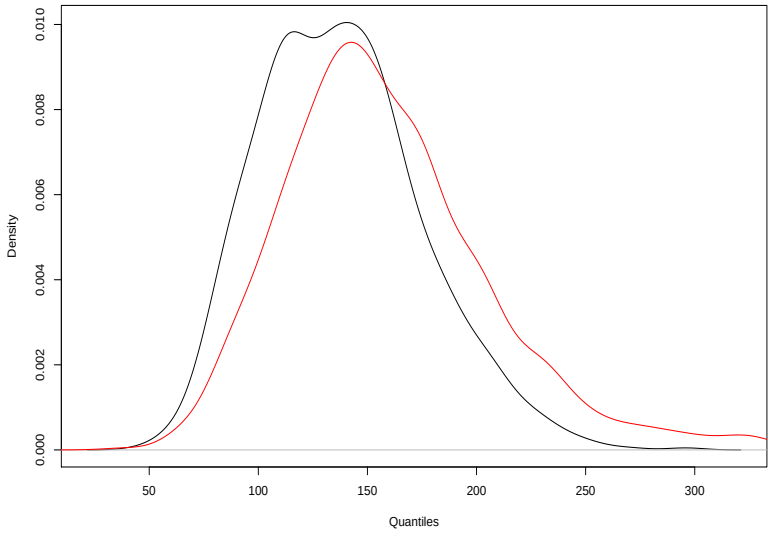

Figure 27. $d = 2$, Mixture Normal, $n = 100$, Kernel Densities

Now we examine the Q-Q plot for $n = 1000$ in Figure 28. This is followed by the kernel densities for $n = 1000$ in Figure 29 and the Q-Q plot for $n = 5000$ in Figure 30.

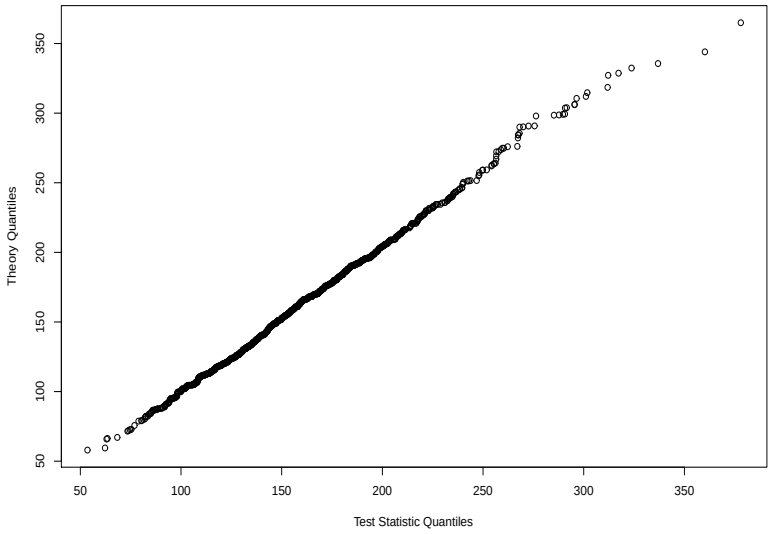

Figure 28. $d = 2$, Mixture Normal, $n = 1000$, Q-Q Plot

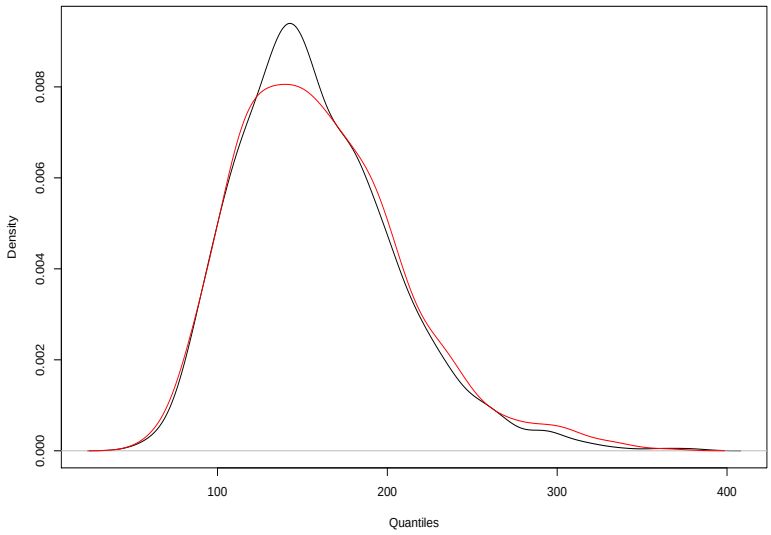

Figure 29. $d = 2$, Mixture Normal, $n = 1000$, Kernel Densities

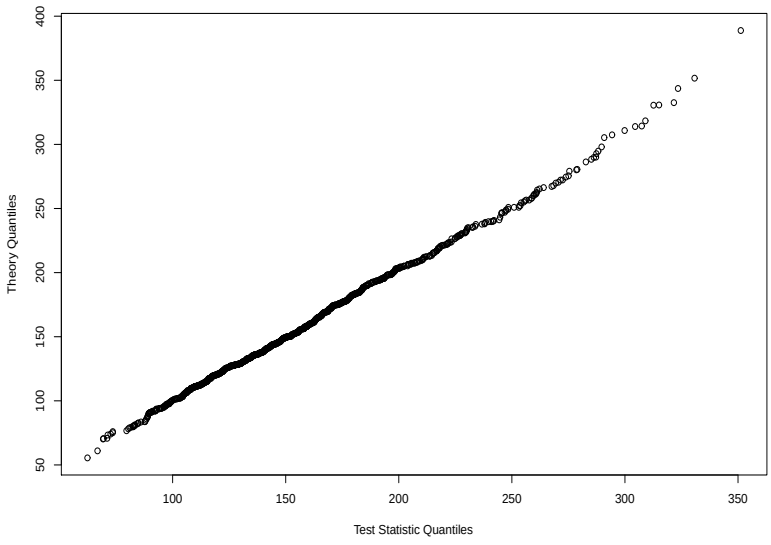

Figure 30. $d = 2$, Mixture Normal, $n = 5000$, Q-Q Plot

Finally we examine the kernel densities for $n = 5000$ in Figure 31.

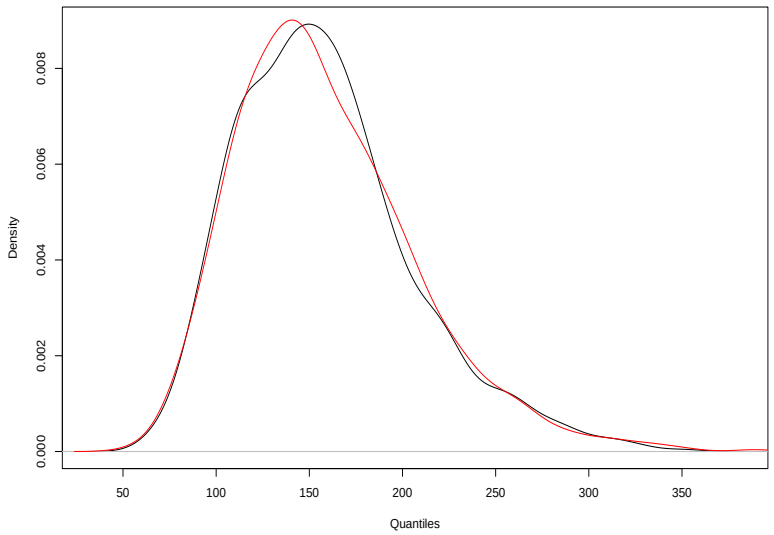

Figure 31. $d = 2$, Mixture Normal, $n = 5000$, Kernel Densities

For our last setting, we add covariance differences to group 2, making the new dimension $d = 4$. The new covariance matrices for group 2 are the following.

$$\mathbf{\Sigma_{421}} = \mathbf{\Sigma_{221}} + \begin{bmatrix} +0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & +0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & +0 & 0 & 0 & 0 & 0 & 1.\bar{2} \\ 0 & 0 & 0 & +0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & +0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & +0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & +0 & 0 \\ 0 & 0 & 1.\bar{2} & -1 & 0 & 0 & 0 & +0 \end{bmatrix} \tag{6.51}$$

$$\boldsymbol{\Sigma_{422}} = \boldsymbol{\Sigma_{221}} + \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & .\bar{2} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & .\bar{2} & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{6.52}$$

$$\boldsymbol{\Sigma_{42}} = \boldsymbol{\Sigma_{22}} + \begin{bmatrix} +0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & +0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & +0 & 0 & 0 & 0 & 0 & .7\bar{2} \\ 0 & 0 & 0 & +0 & 0 & 0 & 0 & -.5 \\ 0 & 0 & 0 & 0 & +0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & +0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & +0 & 0 \\ 0 & 0 & .7\bar{2} & -.5 & 0 & 0 & +0 & 0 \end{bmatrix} \tag{6.53}$$

The other parameters remain the same.

$$
\begin{aligned}
\boldsymbol{\mu_{411}} &= \boldsymbol{\mu_{211}} \\
\boldsymbol{\mu_{412}} &= \boldsymbol{\mu_{212}} \\
\boldsymbol{\mu_{421}} &= \boldsymbol{\mu_{221}} \\
\boldsymbol{\mu_{422}} &= \boldsymbol{\mu_{222}} \\
\boldsymbol{\mu_{431}} &= \boldsymbol{\mu_{231}} \\
\boldsymbol{\mu_{432}} &= \boldsymbol{\mu_{232}} \\
\boldsymbol{\Sigma_{411}} &= \boldsymbol{\mu_{211}} \\
\boldsymbol{\Sigma_{412}} &= \boldsymbol{\mu_{212}} \\
\boldsymbol{\Sigma_{431}} &= \boldsymbol{\mu_{231}} \\
\boldsymbol{\Sigma_{432}} &= \boldsymbol{\mu_{232}} \\
\boldsymbol{\mu_{41}} &= \boldsymbol{\mu_{21}} \\
\boldsymbol{\mu_{42}} &= \boldsymbol{\mu_{22}} \\
\boldsymbol{\mu_{43}} &= \boldsymbol{\mu_{23}} \\
\boldsymbol{\Sigma_{41}} &= \boldsymbol{\Sigma_{21}} \\
\boldsymbol{\Sigma_{43}} &= \boldsymbol{\Sigma_{23}}
\end{aligned}
\tag{6.54}
$$

The eigenvalues of this $d = 4$ situation are the following.

$$
\begin{aligned}
\lambda_{41} &= 1.666935 \\
\lambda_{42} &= 0.801735 \\
\lambda_{43} &= .02851263 \\
\lambda_{44} &= .01381289 \\
\lambda_{45} &= \ldots = \lambda_{48} = 0
\end{aligned}
\tag{6.55}
$$

We performed similar simulations to those made for the $d = 2$. We tried an additional high sample size of $n = 10000$. In this complex case, good convergence only happened once we reached this large size.

We begin examination with the Q-Q plot for $n = 100$ in Figure 32. Then we look at the kernel densities for $n = 100$ in Figure 33 and the Q-Q plot for $n = 1000$ in Figure 34.

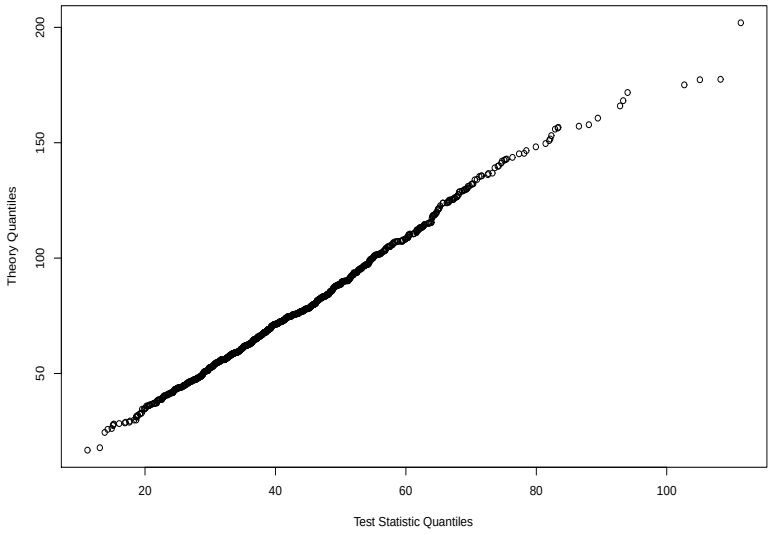

Figure 32. $d = 4$, Mixture Normal, $n = 100$, Q-Q Plot

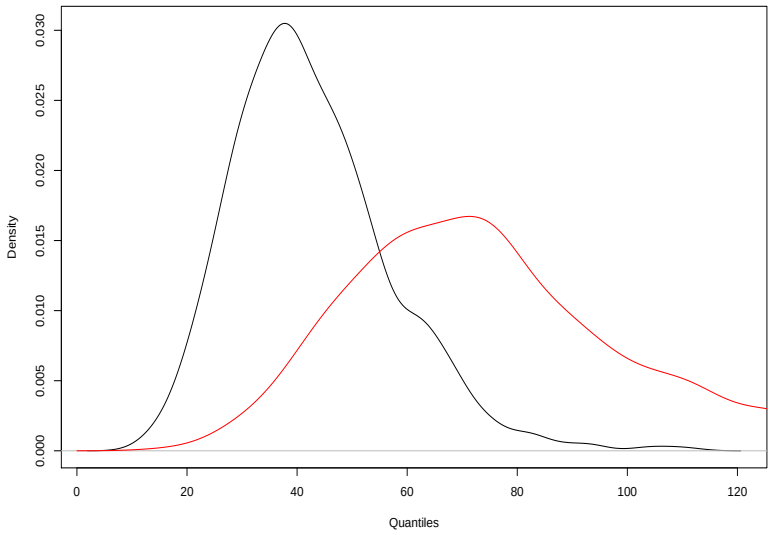

Figure 33. $d = 4$, Mixture Normal, $n = 100$, Kernel Densities

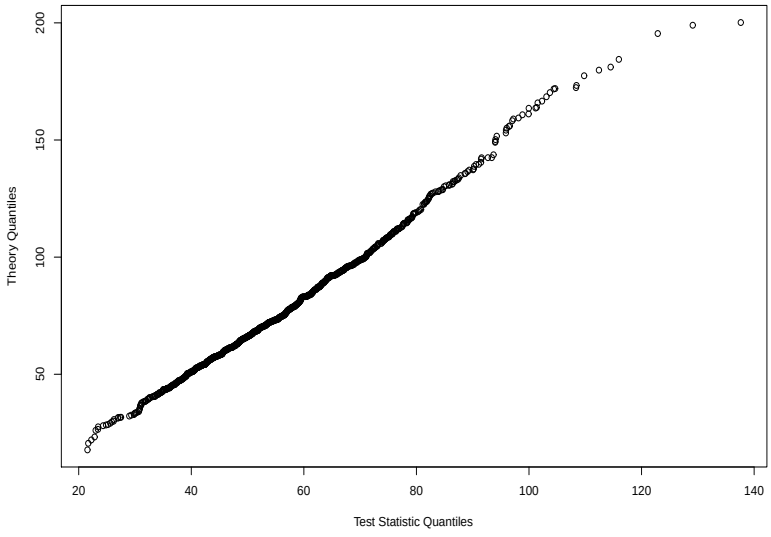

Figure 34. $d = 4$, Mixture Normal, $n = 1000$, Q-Q Plot

Now we examine the kernel densities for $n = 1000$ in Figure 35. This is followed by the Q-Q plot for $n = 5000$ in Figure 36 and the kernel densities for $n = 5000$ in Figure 37.

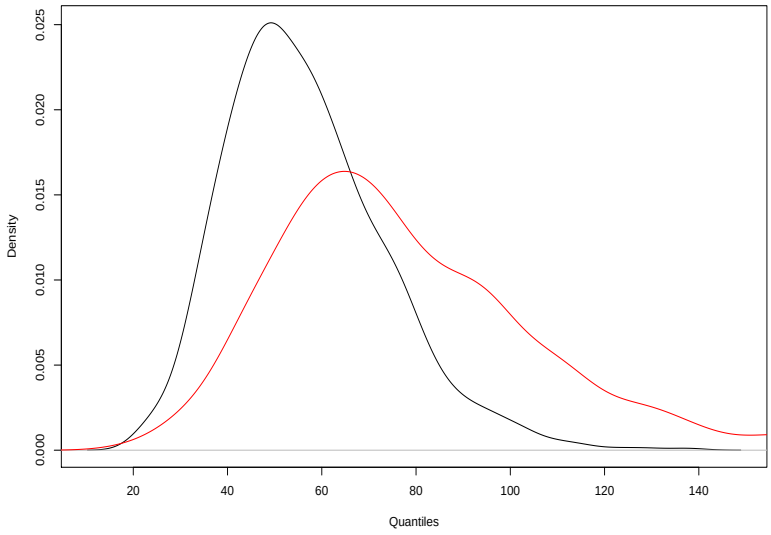

Figure 35. $d = 4$, Mixture Normal, $n = 1000$, Kernel Densities

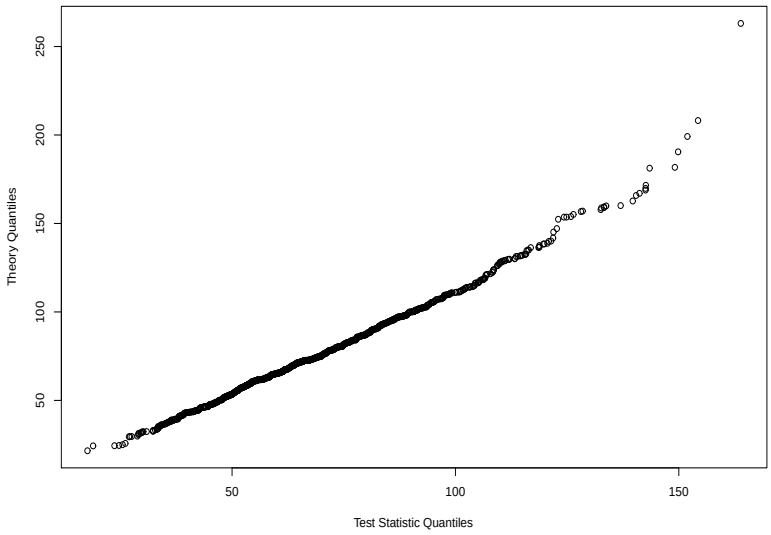

Figure 36. $d = 4$, Mixture Normal, $n = 5000$, Q-Q Plot

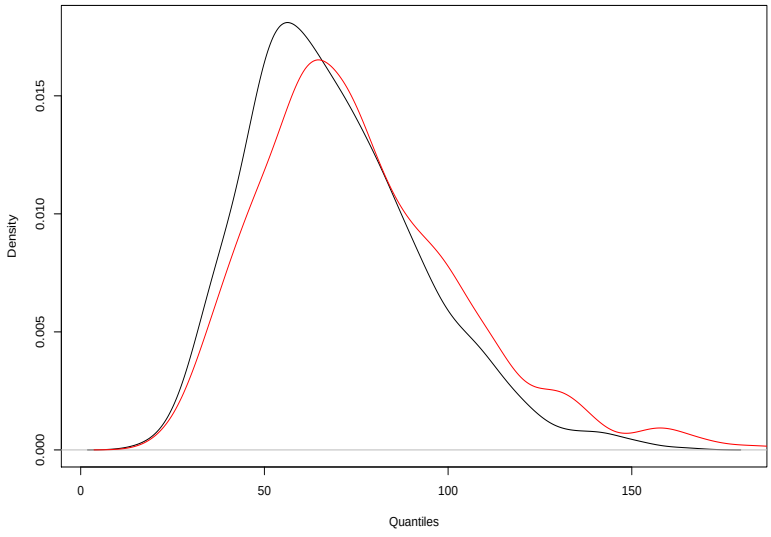

Figure 37. $d = 4$, Mixture Normal, $n = 5000$, Kernel Densities

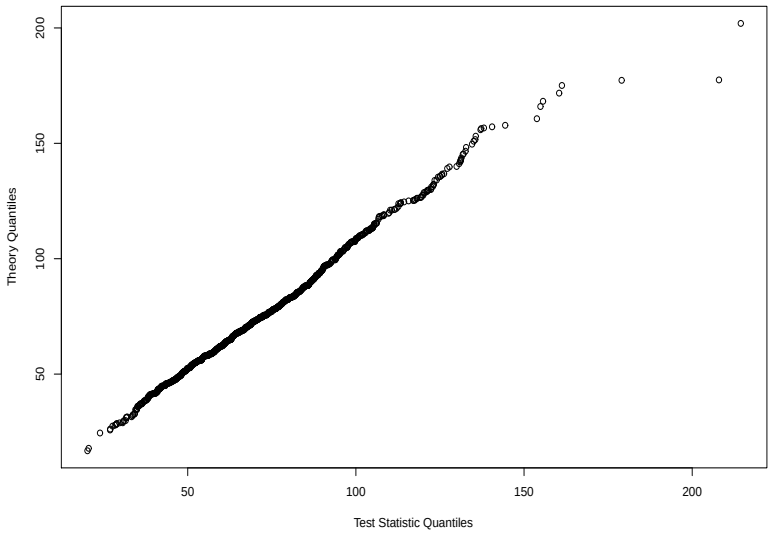

Figure 38. $d = 4$, Mixture Normal, $n = 10000$, Q-Q Plot

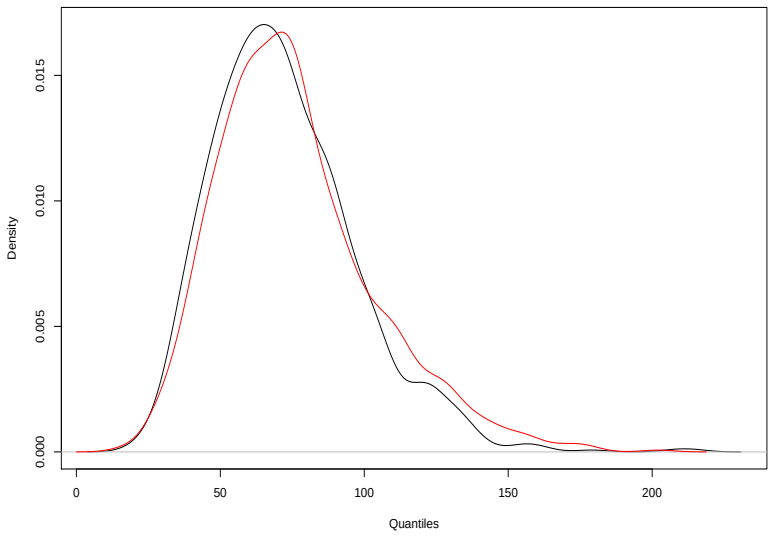

Figure 39. $d = 4$, Mixture Normal, $n = 10000$, Kernel Densities

We finally see results that suggest good convergence at $n = 10000$. The Q-Q plot for this sample size is given in Figure 38. The kernel densities for $n = 10000$ are in Figure 39.

2. Multivariate Normal

We also tested each of the last 3 situations under a multivariate normal population setting. The marginal group parameters were used to parameter the multivariate normal populations. The eigenvalue and dimensionality results remain the same, as they are calculated using only the marginal group parameters.

In the $d = 1$ case, the theory (4.27) reference distribution matched that of the test statistic for the higher sample sizes. This is similar to the mixture setting. We begin by examining the Q-Q plot for $n = 100$ in Figure 40. Then we examine the kernel densities for $n = 100$ in Figure 41 and the Q-Q Plot for $n = 1000$ in Figure 42.

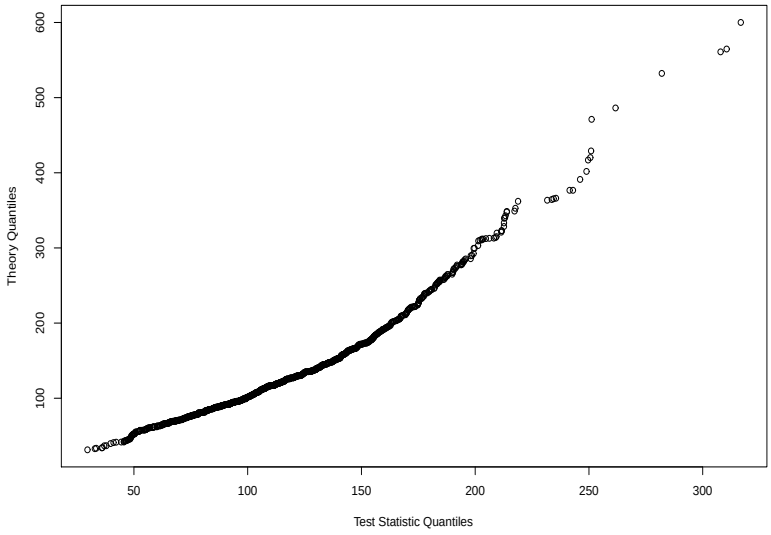

Figure 40. $d = 1$, Normal, $n = 100$, Q-Q Plot

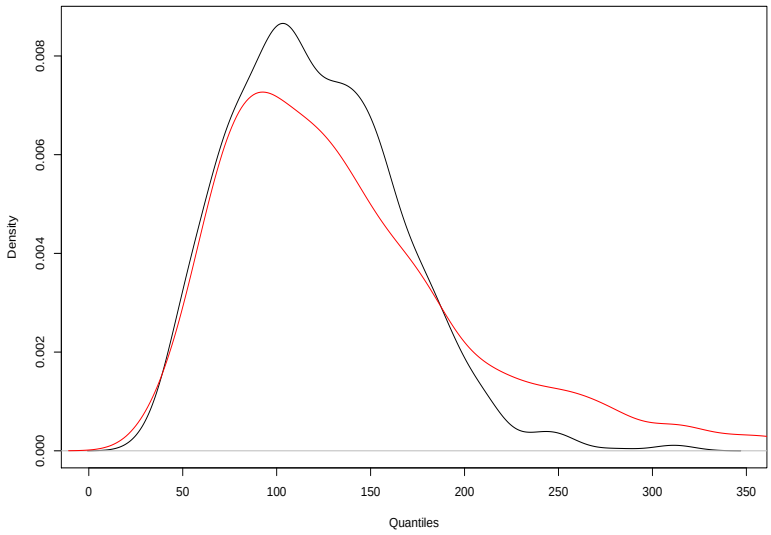

Figure 41. $d = 1$, Normal, $n = 100$, Kernel Densities

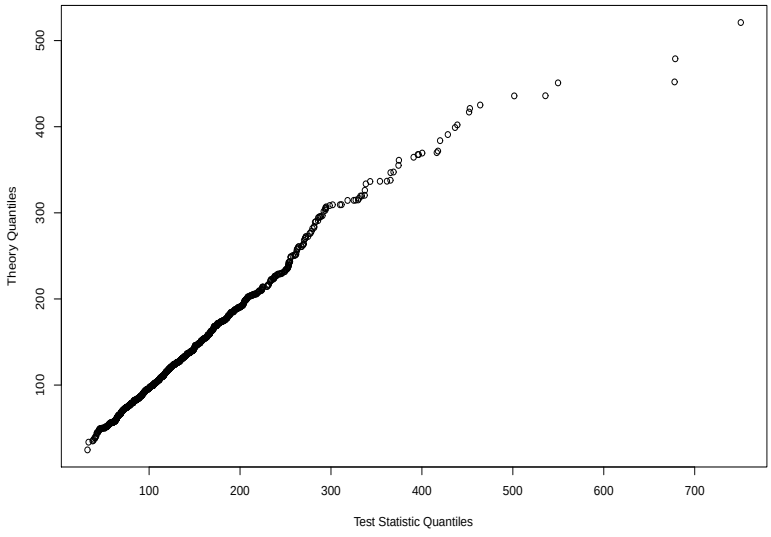

Figure 42. $d = 1$, Normal, $n = 1000$, Q-Q Plot

Now we examine the kernel densities for $n = 1000$ in Figure 43. We follow by rendering the Q-Q Plot for $n = 5000$ in Figure 44 and the kernel densities for $n = 5000$ in Figure 45.

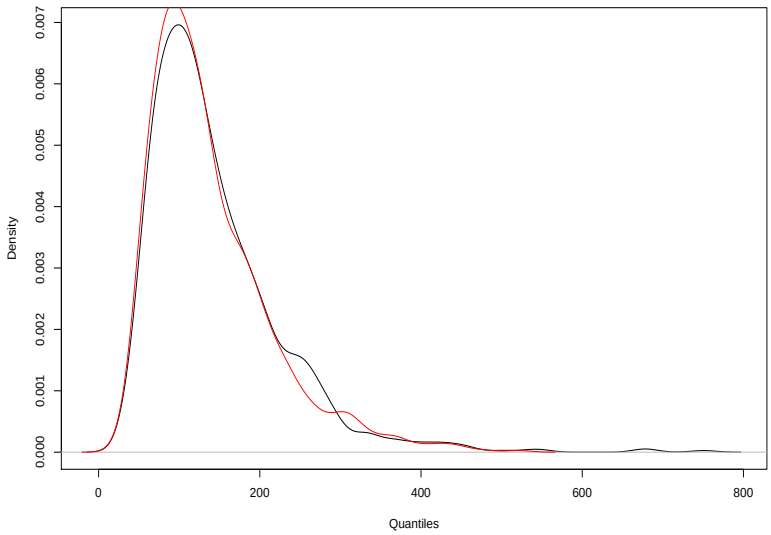

Figure 43. $d = 1$, Normal, $n = 1000$, Kernel Densities

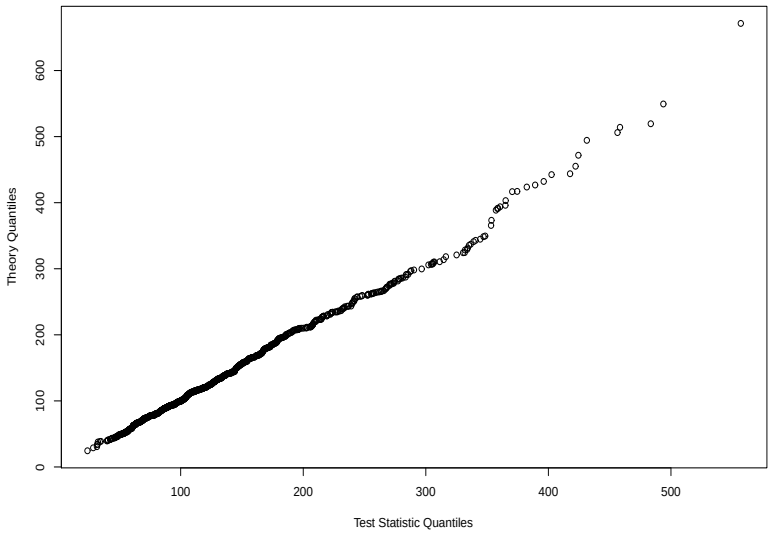

Figure 44. $d = 1$, Normal, $n = 5000$, Q-Q Plot

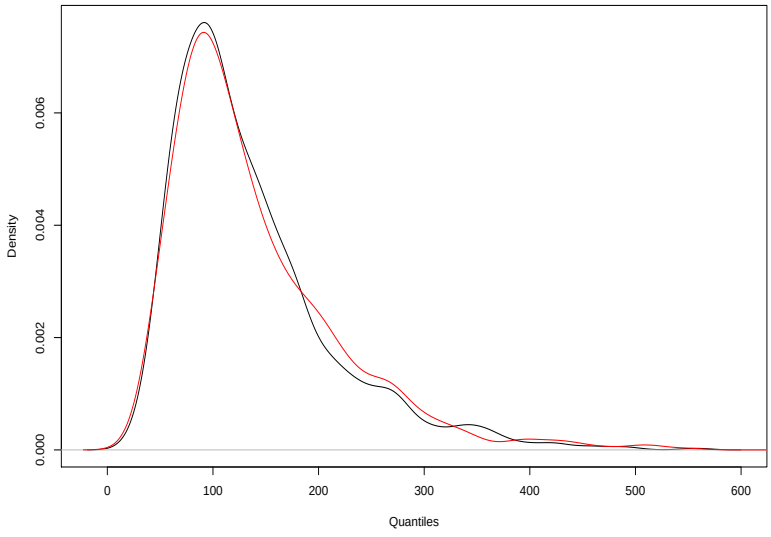

Figure 45. $d = 1$, Normal, $n = 5000$, Kernel Densities

For the $d = 2$ case, we also see good convergence results for $n = 1000$ and above. The convergence is clearly better when $n = 1000$ under the normal population case than the mixture case. We begin by examining the Q-Q plot for $n = 100$ in Figure 46. This is followed by the kernel densities for $n = 100$ in Figure 47 and the Q-Q plot for $n = 1000$ in Figure 48.

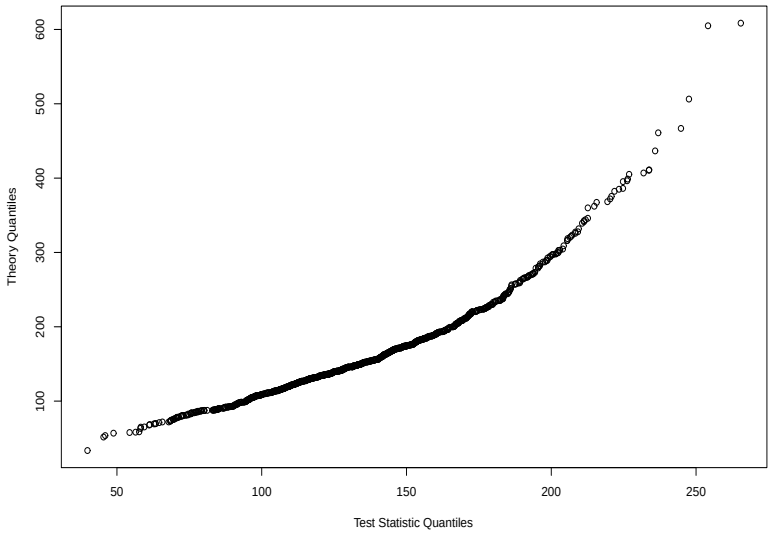

Figure 46. $d = 2$, Normal, $n = 100$, Q-Q Plot

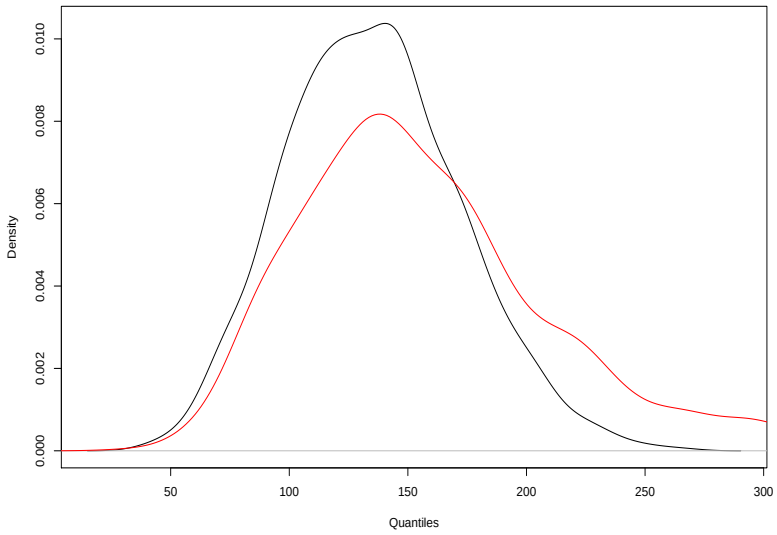

Figure 47. $d = 2$, Normal, $n = 100$, Kernel Densities

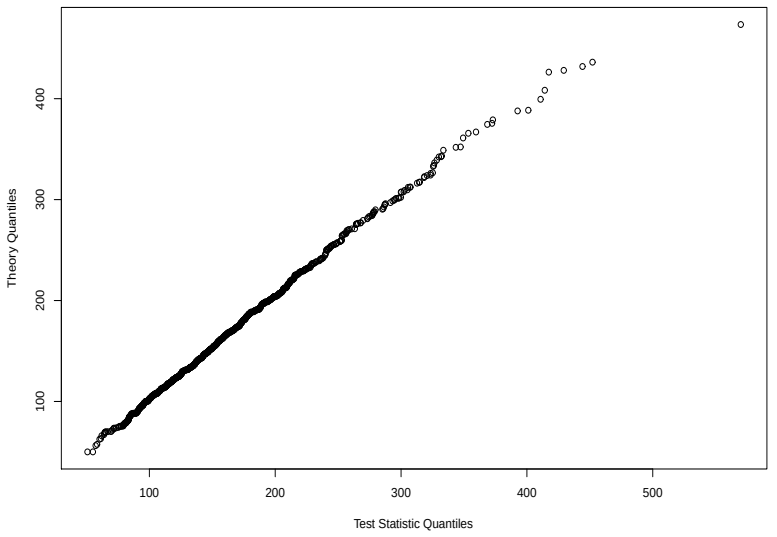

Figure 48. $d = 2$, Normal, $n = 1000$, Q-Q Plot

Next we examine the kernel density plots for $n = 1000$ in Figure 49. The Q-Q plot for $n = 5000$ follows in Figure 50 along with the kernel density plots for $n = 5000$ in Figure 51.

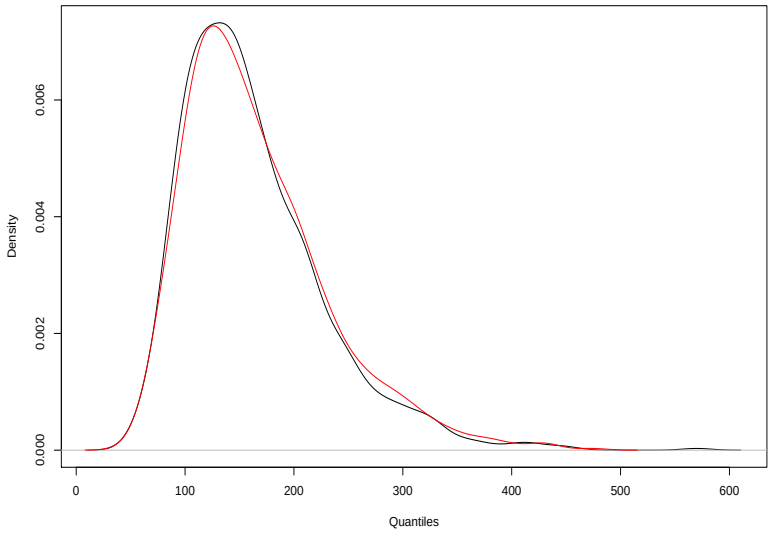

Figure 49. $d = 2$, Normal, $n = 1000$, Kernel Densities

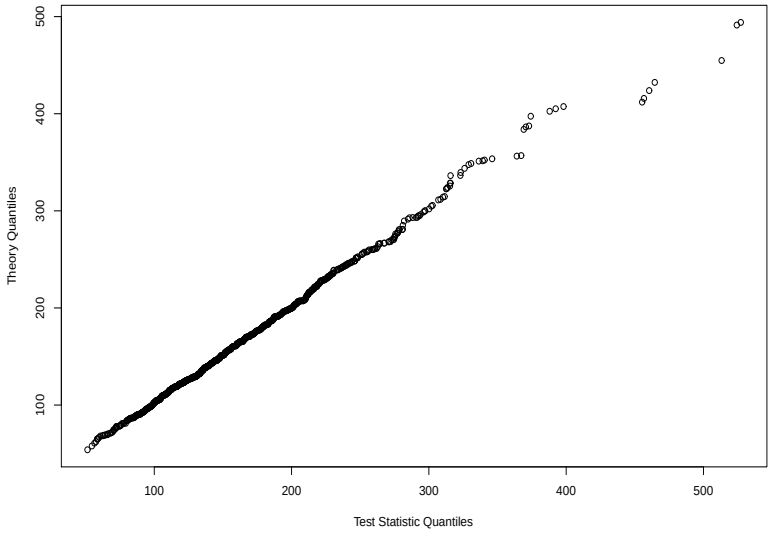

Figure 50. $d = 2$, Normal, $n = 5000$, Q-Q Plot

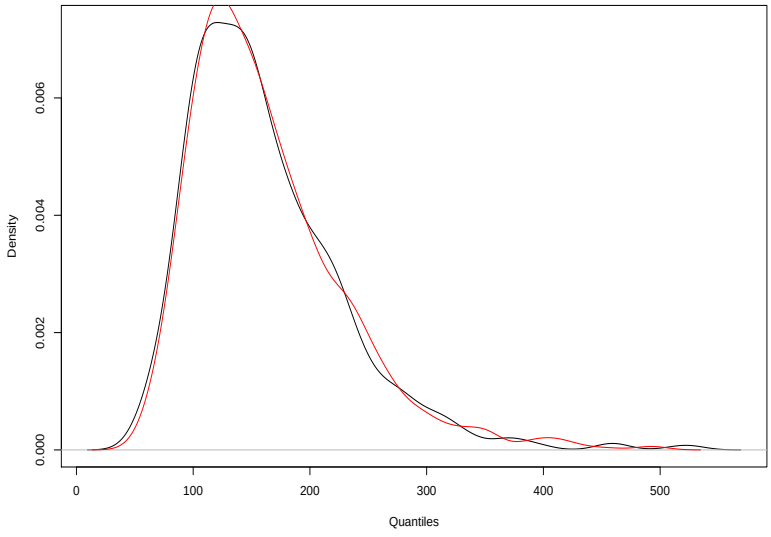

Figure 51. $d = 2$, Normal, $n = 5000$, Kernel Densities

For $d = 4$, like the mixture case the normal setting only reaches adequate convergence once $n = 10000$. Interestingly, the theory and actual distributions match better under the mixture case at $n = 10000$ than in the normal case. We begin with the Q-Q plot for $n = 100$ in Figure 52. This is followed by the kernel density plots at $n = 100$ in Figure 53 and the Q-Q plot for $n = 1000$ in Figure 54.

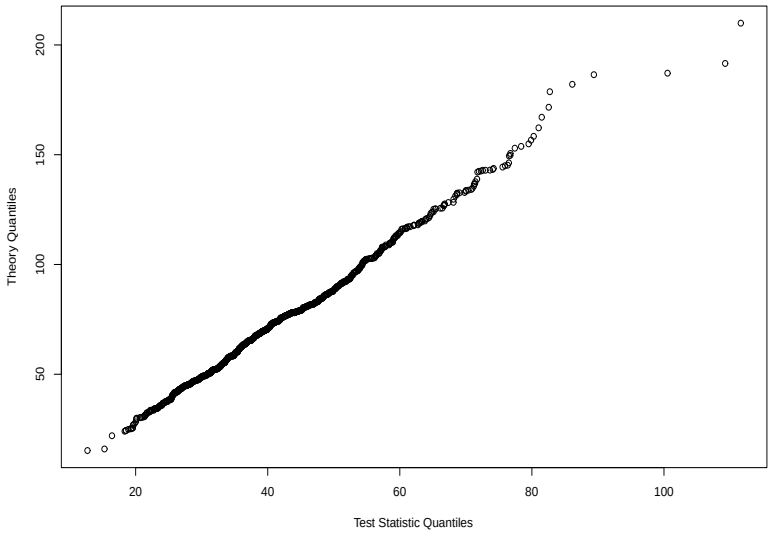

Figure 52. $d = 4$, Normal, $n = 100$, Q-Q Plot

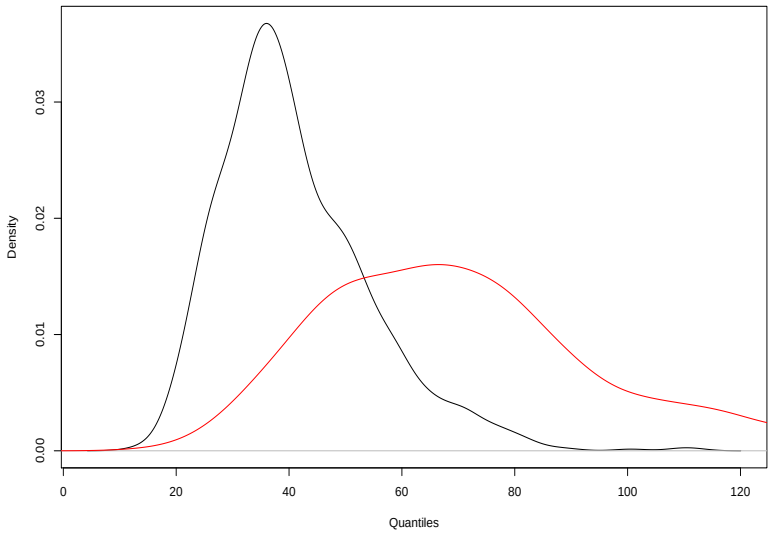

Figure 53. $d = 4$, Normal, $n = 100$, Kernel Densities

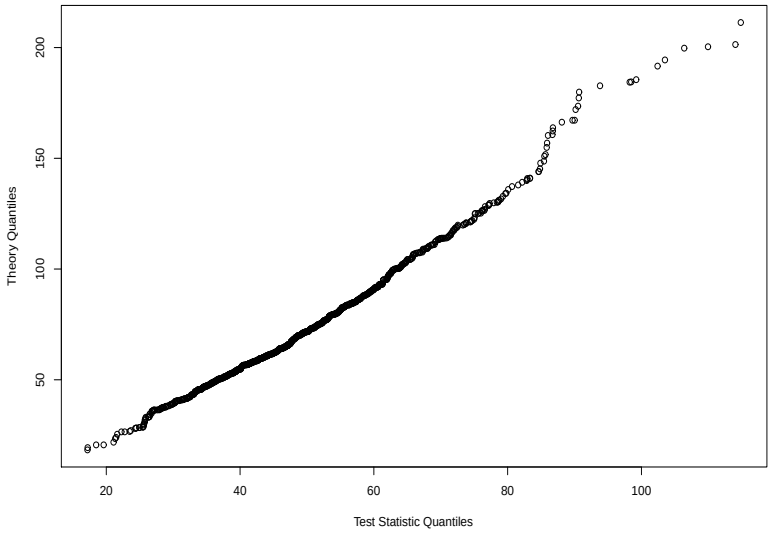

Figure 54. $d = 4$, Normal, $n = 1000$, Q-Q Plot

Next we examine the kernel densities for $n = 1000$ in Figure 55. This is followed by the Q-Q plot for $n = 5000$ in Figure 56 and the kernel density plots for $n = 5000$ in Figure 57.

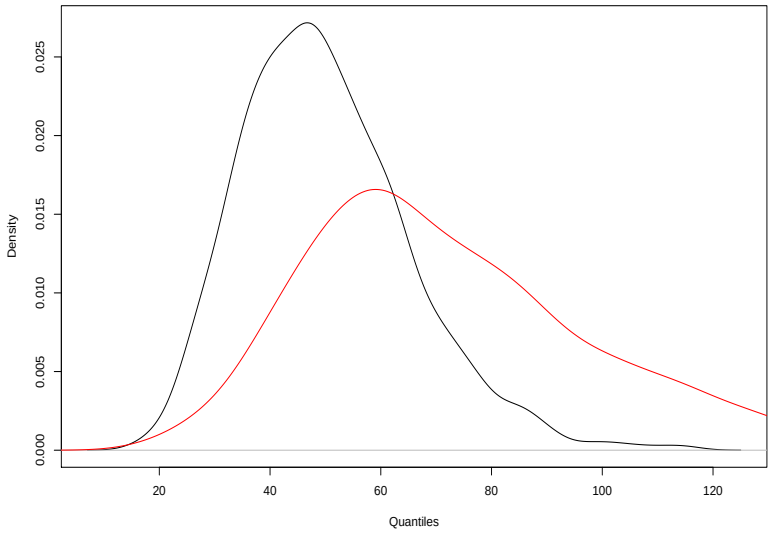

Figure 55. $d = 4$, Normal, $n = 1000$, Kernel Densities

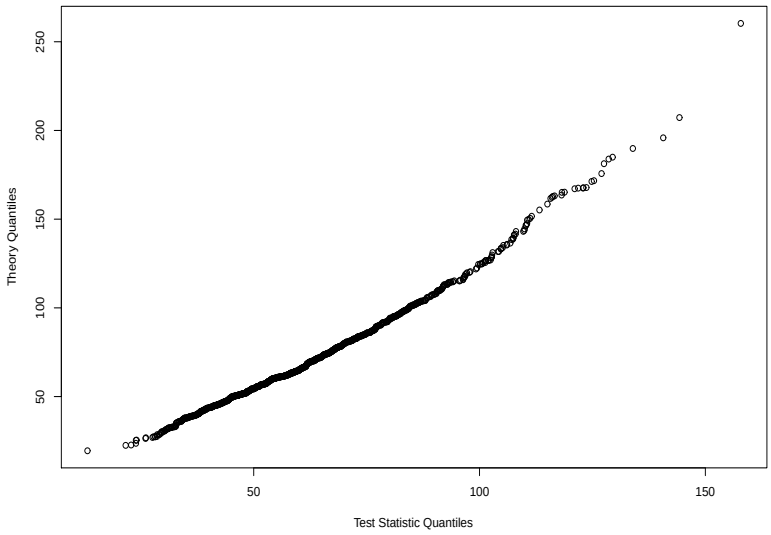

Figure 56. $d = 4$, Normal, $n = 5000$, Q-Q Plot

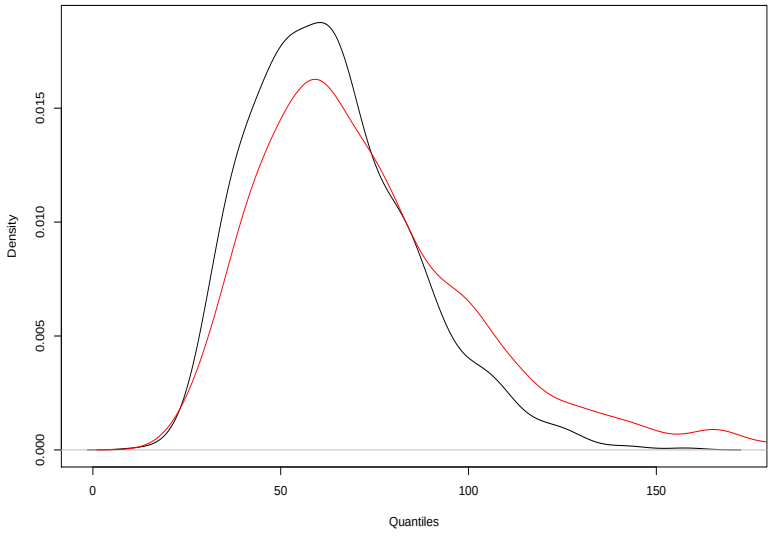

Figure 57. $d = 4$, Normal, $n = 5000$, Kernel Densities

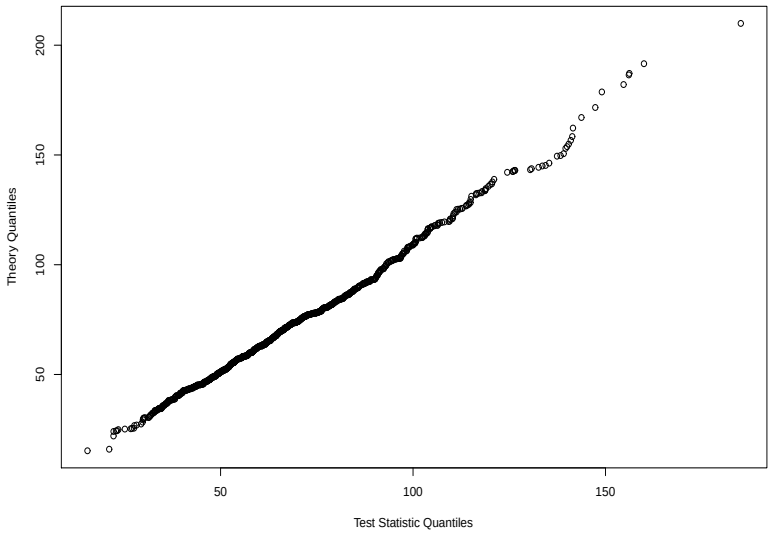

Figure 58. $d = 4$, Normal, $n = 10000$, Q-Q Plot

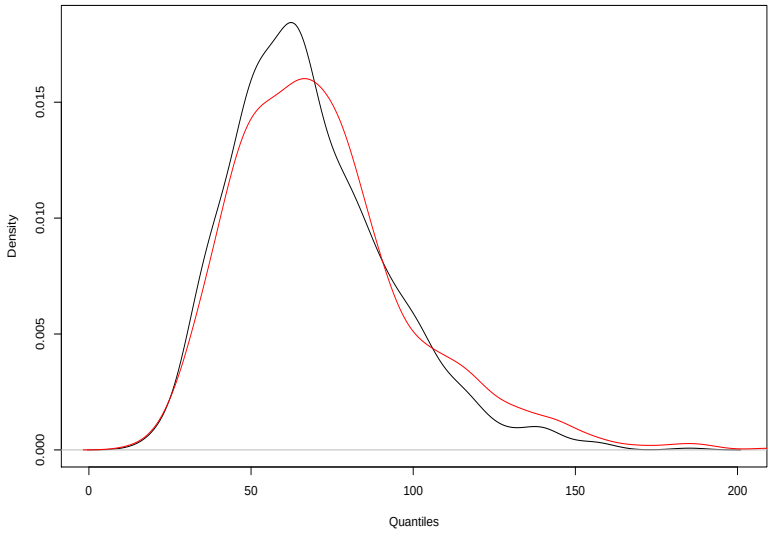

Figure 59. $d = 4$, Normal, $n = 10000$, Kernel Densities

We end our investigation of the normal setting with the Q-Q Plot for $n = 10000$ in Figure 58 and the kernel density plots for $n = 10000$ in Figure 59. Comparison of kernel density matching in Figure 59 with the mixture normal analog in Figure 39 shows how the mixture case achieves better convergence at $n = 10000$.

In this section, we have used multivariate normal and mixture multivariate normal population data collected under 3 varied SMVCIR dimension settings. Our results have been good, showing that the theory (4.27) reference distribution and the actual distribution are equivalent in all the examined settings. What deviations we have seen are minor, and had little effect on the tails of the distributions. The tails are essential for our tests, as the empirical (4.28) and approximate empirical (4.30) tests approximate the tails of the theory (4.27) distribution to form a decision rule.

This concludes our simulation studies. We have demonstrated the proper behavior of the SMVCIR dimensionality test methods: the iterated testing scheme of (4.1), the theory (4.27) null reference distribution, the empirical (4.28) null reference distribution, and finally the approximate empirical (4.28) null reference distribution. A variety of different situations were considered, and the SMVCIR methods behave well in each.

CHAPTER VII

REAL DATA EXAMPLES

In this chapter we will apply SMVCIR to some real world examples. We first revisit the pen digit example from Sheather et al. (2008). This data was also previously studied in Zhu and Hastie (2003).

The pen digit database reference in Zhu and Hastie (2003) contains samples of the handwritten digits 0-9 from 44 separate writers. Data on 16 variables was recorded in each sample. The data was divided into a training and learning dataset. We will use the learning dataset here. We further restrict our attention to the digits 0 and 9, as these were the most different on the recorded variables. There are a total of 1499 observations in the data.

The pen digit data is also referenced in the new multivariate statistics textbook Izenman (2008). In Chapter VIII of this text, Izenman discusses linear and quadratic discriminant analysis (lda and qda). These techniques create a decision rule to discriminate between groups. The inverse regression methods we have studied here SIR, SAVE, and SMVCIR provide mainly graphical group discrimination.

There is a similarity, if not equivalence in some cases, between the decision rules used in lda and qda and the graphical results from SIR, SAVE, and SMVCIR. QDA separates the group based on their group covariances. LDA discriminates between groups by using the group means. One of the examples, "wine" contains information over 13 variables for 178 wines grown in Italy during 1970-1979. Both qda and lda perform well on this dataset. We will study this dataset here, evaluating how SIR, SAVE, and SMVCIR performed on it.

In these examples, the pen digit from Sheather et al. (2008) and wine from Izenman (2008), we will see how SMVCIR performs and compares with SAVE and

SIR in several aspects. They will not provide a definitive picture of how each algorithm performs, but they will provide a broad and varied set of real examples. Our simulated examples were useful, but their interpretation may be esoteric to some readers. Real examples should provide better interpretation of our results, as they may be easier for the reader to relate to.

The primary comparison will be between SMVCIR's group discrimination of Mean, Variance, and Covariance and the discrimination provided by SAVE and SIR. In this dissertation we provided an augmentation to the SMVCIR algorithm that uses scaling standardization instead of Mahalanobis standardization. SAVE and SIR still use the Mahalanobis standardization. The previous comparisons of SMVCIR with SAVE and SIR in Sheather et al. (2008) used the Mahalanobis standardization in all three algorithms.

We will also compare the dimensionality test results for each of the three algorithms. The empirical (4.28) and approximate empirical (4.30) will be compared as well.

A. Pen Digit

We input the Pen Digit data into SMVCIR. First we obtain the dimensionality test results using the iterated testing scheme (4.1). The results are shown in Table 61. We choose to use $\alpha = .05$ as our individual test size. We finally accept the null hypothesis when $d_0 = 14$.

Table 61. Pen Digit, SMVCIR d Test

Dimension	Ap. Emp.	Emp.
0	0	0
⋮	⋮	⋮
11	0	0
12	0	0.001
13	0.0198	0.016
14	0.4978	0.469

The eigenvalues of the estimated SMVCIR kernel are given below.

$$\begin{aligned}
\lambda_1 &= 5.808416 \\
\lambda_2 &= 5.553553 \\
\lambda_3 &= 2.434885 \\
\lambda_4 &= 1.140898 \\
\lambda_5 &= 0.5212225 \\
\lambda_6 &= 0.2032391 \\
\lambda_7 &= 0.09856586 \\
\lambda_8 &= 0.08070066 \\
\lambda_9 &= 0.04495229 \\
\lambda_{10} &= 0.03461531 \\
\lambda_{11} &= 0.01897307 \\
\lambda_{12} &= 0.01435828 \\
\lambda_{13} &= 0.009682227 \\
\lambda_{14} &= 0.003913212 \\
\lambda_{15} &= 0.0007299168 \\
\lambda_{16} &= 0.00009676446
\end{aligned} \tag{7.1}$$

We see that the first 5 dimensions have appreciable magnitude. The remaining nine statistically significant dimensions are negligible. We have a large sample size, 1499 observations, so we can use the dimensionality test results with confidence.

For our purposes here, we will not go into an exceptionally detailed analysis of the pen digit data. So we will only examine the first five highest discriminating dimensions. We acknowledge that there is evidence that the remaining 8 dimensions provide accurate discrimination between the 0 and 9 digit groups, but their discrimination is not practically significant for us here.

In Figure 60, we produced a scree plot of the singular values of the SMVCIR spanning set to justify our practical significance decision. Recall that the eigenvalues of the SMVCIR kernel represent the discrimination strength of the SMVCIR dimensions. These singular values are the square roots of the eigenvalues, so they represent discrimination strength as well. We see that nearly 80% of the total discrimination power (as measured by the summed singular values) is represented in the first five SMVCIR dimensions.

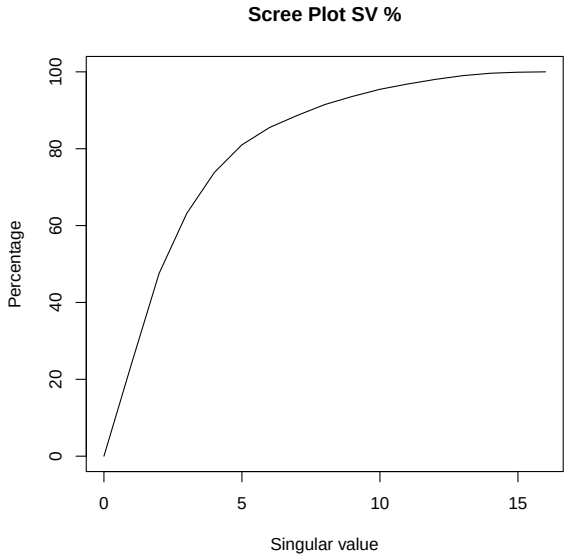

Figure 60. Pen Digit, SMVCIR Singular Values Scree Plot

Now we will examine the SMVCIR coordinates in Figure 61. The 0 group is black while 9 group is red. Interpretation of whether each dimension corresponds to a variance, mean, or covariance difference is not totally clear. As we have discussed before, we are seeing the underlying latent dimensions of the SMVCIR space. These dimensions may be made of several of the spanning set vector differences. Also, the scaling standardization that we use here is only appropriate when the resulting coordinates are not highly correlated. None of the SMVCIR coordinates are highly correlated in this case.

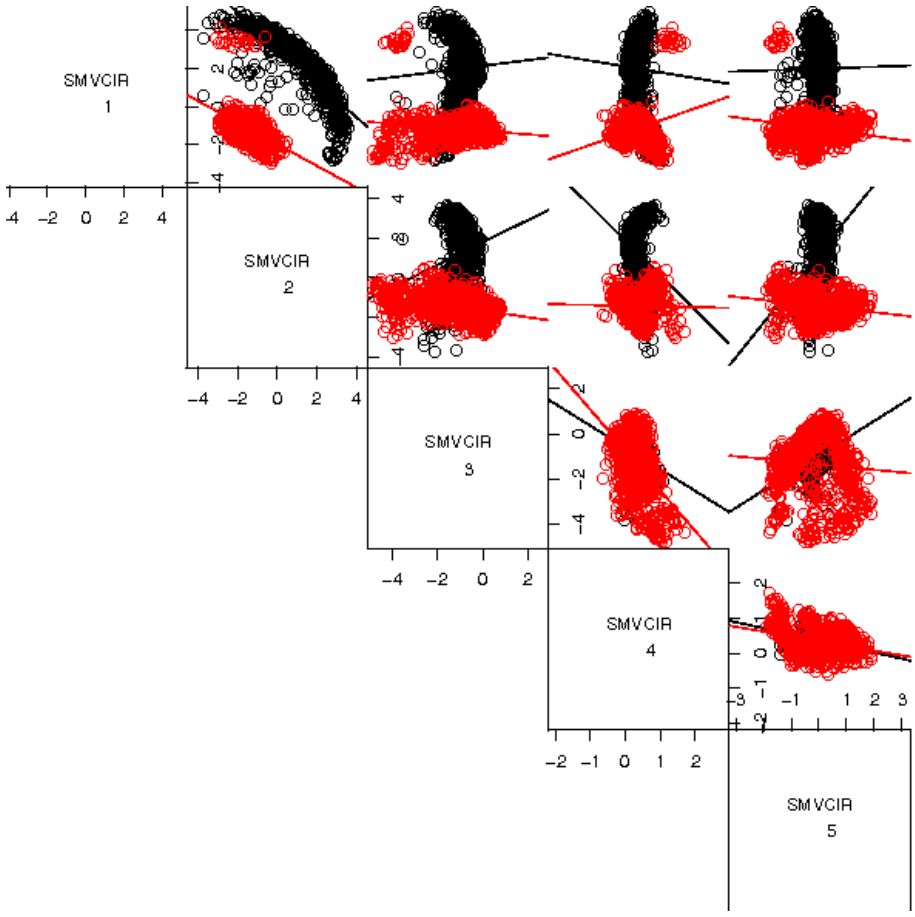

Figure 61. Pen Digit, SMVCIR Coordinates

Each of the dimensions, especially the first two show a sharp variance difference between the groups. The first dimension also provides a location difference, and a clean clustering of the 9 group into two subgroups. This is interesting. Examination of these clusters may yield useful insight into the 9 group. The remaining groups show covariance differences.

We also look at a 3D plot using rgl package Adler and Murdoch (2009). This plot is given in Figure 62.

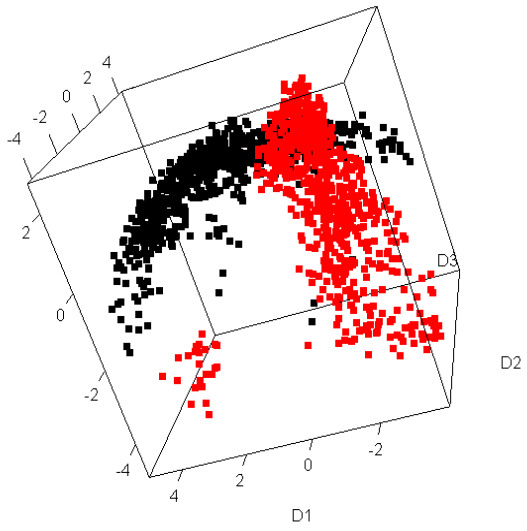

Figure 62. Pen Digit, SMVCIR Coordinates D1-D3

The power of SMVCIR to discriminate between the groups appears to have increased under the new standardization method. The types of differences found in Sheather et al. (2008) coincide with what we see here, but the scaling discrimination has increased. The 2 subgroup point clustering in group 9 was not detected under the old standardization method as well.

We regress the first three SMVCIR dimensions on the predictors, giving the results in Table 62. Standardized coefficients, which add up to 1 when squared, are reported as in Sheather et al. (2008). The first dimension puts significant negative weight on predictors x_1, x_4, x_9, and x_{13}. Significant weight, in the opposite direction is placed on x_{12}. The second dimension put more balanced weight on the predictors. In the positive direction, x_{13}, x_{14}, and x_{16} are significantly weighted. The predictors x_5, x_7, x_8 and x_{10} are given significant negative weight. The third dimension places significant positive weight on predictors x_1, x_2, and x_9. Significant negative weight is placed on the predictors x_{12} and x_{13} in the third dimension.

Table 62. Pen Digit, SMVCIR Standardized Coefficients

predictor	D_1	D_2	D_3
x_1	-0.307	-0.044	0.284
x_2	-0.107	0.152	0.322
x_3	-0.231	-0.183	-0.089
x_4	-0.353	0.053	0.197
x_5	-0.013	-0.304	-0.158
x_6	-0.372	-0.126	-0.132
x_7	0.209	-0.270	0.339
x_8	-0.250	-0.314	-0.097
x_9	0.214	-0.145	0.480
x_{10}	0.006	-0.436	-0.205
x_{11}	-0.173	0.187	0.091
x_{12}	0.431	-0.098	-0.391
x_{13}	-0.269	0.28	-0.309
x_{14}	0.305	0.275	-0.089
x_{15}	-0.181	0.230	-0.248
x_{16}	0.108	0.443	0.060

Before we analyze the data under the other inverse regression algorithms, we will more thoroughly investigate the two subgroup clustering we discovered within the digit 9 group. We can form the clusters by dividing the digit 9 group into two halves, depending on whether the first SMVCIR coordinate exceeds the value 2 or not.

We find there are only 24 observation in the smaller cluster. Interestingly, each has a zero value for predictor x_1. We draw a kernel density plot of x_1 for values in the larger cluster in Figure 63.

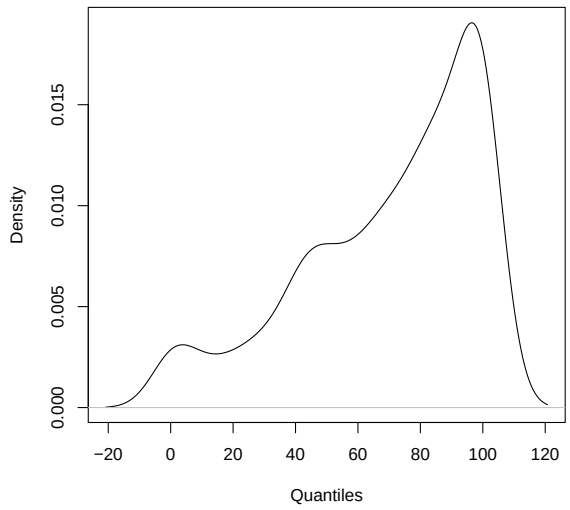

Figure 63. Pen Digit, Digit 9 Large Cluster x_1

There is a clear location difference in x_1 between the two clusters. We use the scaling standardization on the entire digit 9 group and then recalculate the cluster subgroup means to detect other location differences. The results are given in Table 63. We find that the means for the smaller cluster are generally very different from those of the larger. The difference is not unidirectional either. We can classify the points in the small cluster as outliers, and they should be further investigated, especially

on the second and sixteenth predictors (as noted in the table). SMVCIR was able to clearly detect them.

Table 63. Digit 9 Cluster Standardized Means

	Large Cluster	Small Cluster
z_1	0.08	-2.21
z_2	0.14	**-4.01**
z_3	0.02	-0.54
z_4	0.12	-3.37
z_5	-0.03	0.77
z_6	0.11	-3.07
z_7	-0.05	1.4
z_8	0.06	-1.65
z_9	-0.02	0.63
z_{10}	-0.04	1.24
z_{11}	0.07	-2.15
z_{12}	-0.09	2.57
z_{13}	0.04	-1.25
z_{14}	-0.12	3.38
z_{15}	-0.07	1.89
z_{16}	-0.16	**4.67**

When SAVE is applied to the pen digit data, it also finds 14 dimensions when using the iterated testing scheme (4.1) and a .05 test size. We show the first five SAVE dimensions in Figure 64. As reported in Sheather et al. (2008), SAVE reports variance differences for the first five dimensions. Our SMVCIR analysis has the advantage of showing a location difference in the first dimension, and also providing covariance differences in the first five dimensions. The outliers found in the digit 9 clustering that we discovered under SMVCIR are also not seen under SAVE. SIR detects a location difference, and its results are shown in Figure 65. SMVCIR and SAVE both improve upon this by detecting strongly discriminating variance and covariance difference information.

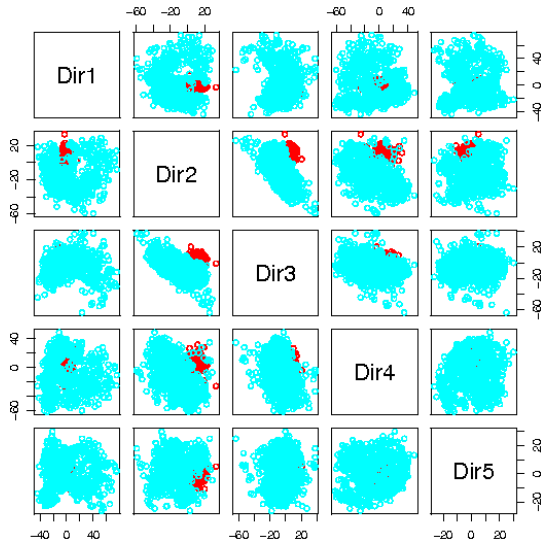

Figure 64. Pen Digit, SAVE Coordinates Dir1-Dir5

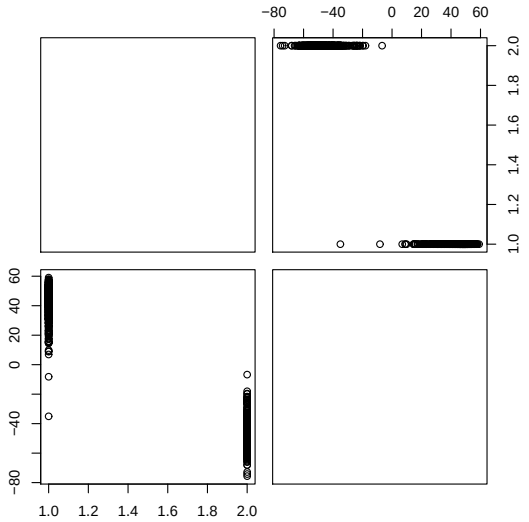

Figure 65. Pen Digit, SIR Coordinate by Digit

B. Wine

The wine dataset from Izenman (2008) records chemical analyses of 178 wines. The wines were all grown in the same part of Italy during 1970-1979. They are classified according to 3 different cultivars: Barbera, Barolo, and Grignolino. There are 59 Barolo wines, 71 Grignolino wines, and 48 Barbera wines.

The chemical attributes recorded were **Alcohol**, **MalicAcid**, **Ash**, **AlcAsh** (Alcalinity of the Ash), **Mg** (Magnesium), **Phenols** (Total Phenols), **Flav** (Flavanoids), **NFP** (Non-Flavanoid Phenols), **Proa** (Proanthocyanims), **Color** (Color Intensity), **Hue**, **OD** (OD280/OD315 of Diluted Wines), and **Proline**.

We invoke SMVCIR on the data, with the dimension test results shown in Table 64. Both the empirical and approximate empirical choose the dimension as $d = 11$.

Table 64. Wine, SMVCIR d Test

Dimension	Ap. Emp.	Emp.
0	0	0
⋮	⋮	⋮
4	0	0
5	0.0001	0.002
6	0.0004	0.002
7	0.0001	0.001
8	0.0002	0
9	0.002	0.005
10	0.0323	0.031
11	0.1113	0.114

A look at the scree plot of singular values of the SMVCIR spanning set estimate (Figure 66) shows us that approximately 70% of the discrimination strength (as measured by the summed singular values) is obtained by using only 5 dimensions. We will not go into an exceptionally detailed analysis of the wine data here. So we will only examine the first five highest discriminating dimensions. We acknowledge that there is evidence that the remaining 6 dimensions provide accurate discrimination between the three groups, but their discrimination is not practically significant for us here.

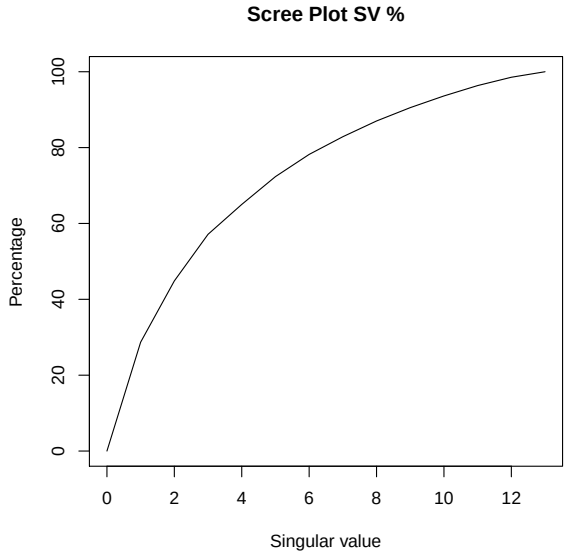

Figure 66. Wine, SMVCIR Singular Values Scree Plot

Now we look at the first five SMVCIR coordinates in Figure 67. We have checked the SMVCIR dimensions are not highly correlated, so the scaling standardization method is accurate to use. Here the black points are Barolo cultivars. The red are Grignolino. The blue points are from the Barbera cultivar.

We see that the first and second SMVCIR dimensions provide a location differences between the three cultivar groups. We can also see variance differences in the

second and third dimensions. The differing covariance between the Barolo and other cultivars can be seen as dimension 1 is plotted against the other coordinates. The difference in covariance between the Barbera and Grignolino cultivar groups can be seen in plots of the second, fourth, and fifth SMVCIR dimensions together.

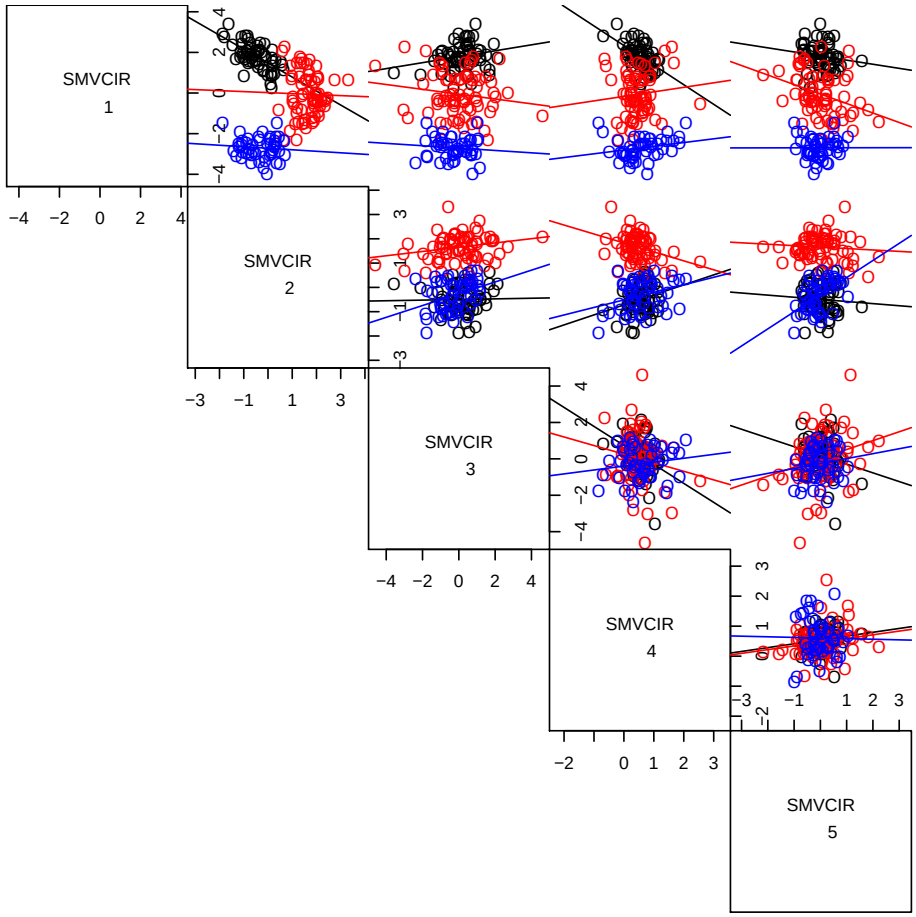

Figure 67. Wine, SMVCIR Coordinates

Table 65 contains the standardized coordinates of the regression of the first SMVCIR coordinate on the 13 chemical analysis attributes.

Table 65. Wine, SMVCIR Standardized Coefficients

predictor	D_1	D_2	D_3	D_4	D_5
Alcohol	0.11	-0.55	0.06	-0.08	0.15
MalicAcid	-0.21	-0.18	-0.32	-0.01	-0.16
Ash	0.05	-0.25	-0.50	0.23	-0.01
AlcAsh	-0.20	0.17	-0.50	0.10	-0.29
Mg	0.12	-0.20	-0.36	0.45	0.44
Phenols	0.38	-0.07	-0.16	-0.33	-0.11
Flav	0.44	0.01	-0.09	0.06	-0.16
NFP	-0.25	-0.01	-0.32	-0.59	0.50
Proa	0.30	0.01	-0.28	-0.31	-0.28
Color	-0.20	-0.53	0.10	-0.32	-0.20
Hue	0.33	0.23	0.00	-0.12	0.50
OD	0.42	0.13	-0.08	-0.16	-0.13
Proline	0.28	-0.44	0.19	0.16	0.03

We begin by examining the location difference in SMVCIR dimension 1. We see high positive mass placed on the predictors **Phenols**, **Flav**, **OD**, **Hue**. Negative mass is placed on the predictors **MalicAcid**, **AlcAsh**, **NFP**, and **Color**.

We see clear location differences in the predictor matrix plot in Figure 68. The predictors with positive coefficients are plotted on the upper left, while the negative coefficient predictors are plotted at the bottom right. Note how the location ordering of the groups between the positive to negative coefficient predictors. This difference in orientation is why the coefficient signs are different.

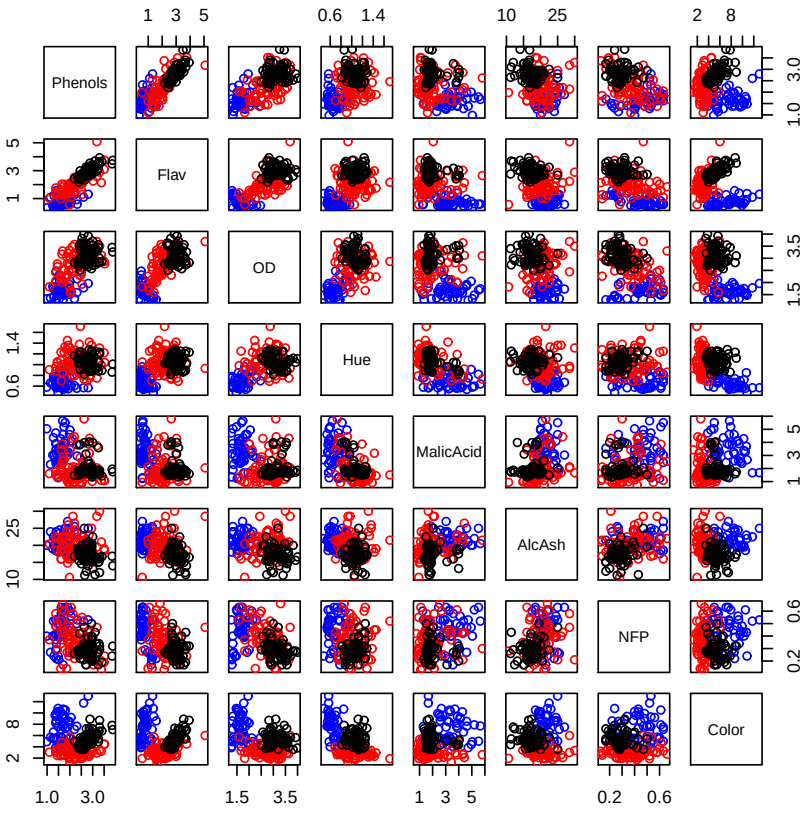

Figure 68. Wine, SMVCIR D_1 Location Differences

For brevity we will omit investigation of the location differences in the second SMVCIR dimension and the small variance differences in the second and third dimensions. For the rest of our study of SMVCIR's output for the wine example, we focus on the covariance differences that SMVCIR indicated.

The sharpest covariance differences involving the first SMVCIR coordinate are found when plotting it versus the second and fourth SMVCIR coordinates. In the standardized coefficients of the first SMVCIR coordinate, particularly high mass is placed on the predictors **Phenols**, **Flav**, and **OD**. In the standardized coefficients for the second SMVCIR coordinates, we see high mass on **Alcohol**, **Color**, and **Proline**.

We draw a matrix plot of the original values for these six predictors in Figure 69. We have added regression lines this time to better show the covariance relationships.

Many of the covariance differences highlighted by comparing the first and second SMVCIR coordinates contrast Barolo (black) versus Grignolino (red) and Barbera (blue). We see this for the covariance of **Proline** with **Alcohol**, **Phenols** and **Flav**. We do see some covariance differences contrasting Grignolino with Barbera and Barolo however. The covariance of **Flav** with **OD** and **Color** have this property.

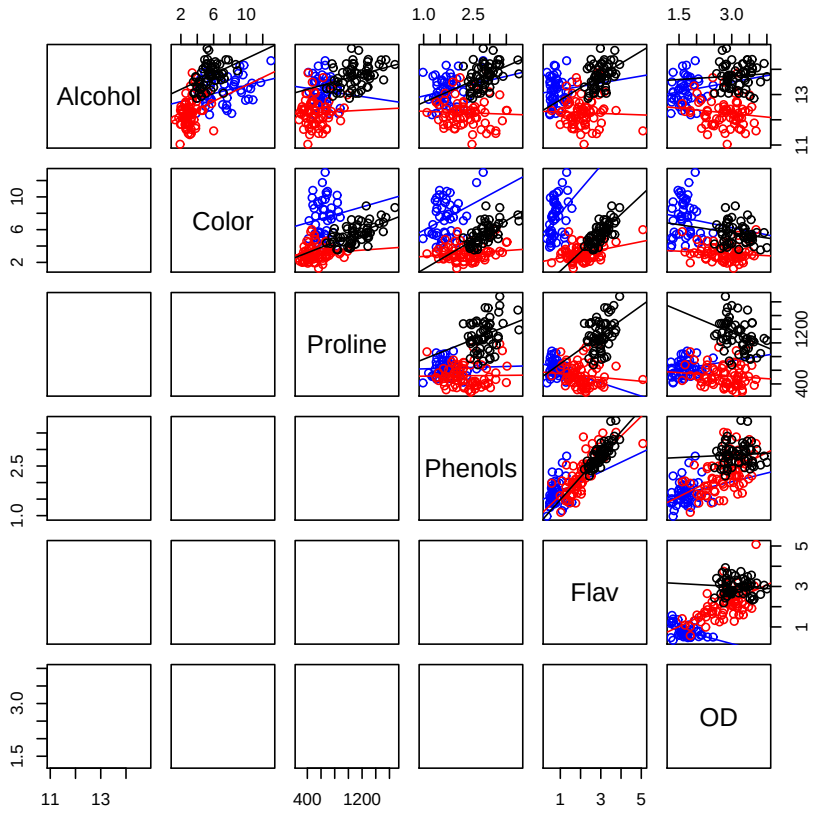

Figure 69. Wine, SMVCIR D_1, D_2 Covariance Differences

The fourth SMVCIR dimension places high mass on the predictor **NFP**. In Figure 70, a matrix plot of this variable with the predictors **Flav** and **Phenols** (which both have high mass coefficients for the first SMVCIR dimension), reveals covariance differences between all three groups for **Phenols** and **NFP**. A covariance difference between Barbera against Grignolino and Barolo is seen for the predictor **Flav** against **Phenols** and **NFP**.

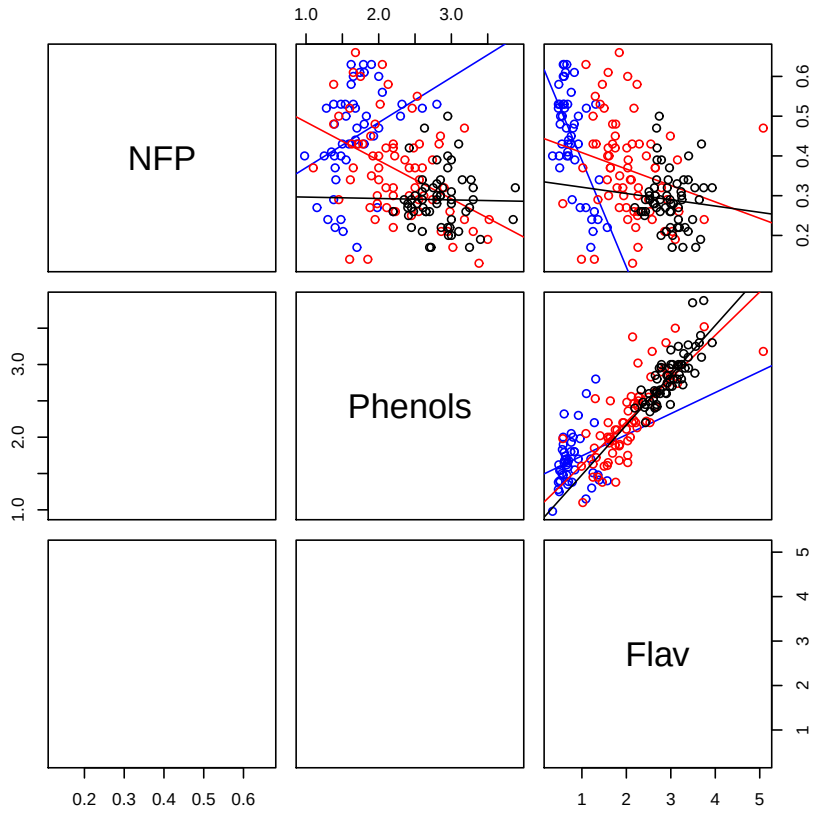

Figure 70. Wine, SMVCIR D_1, D_4 Covariance Differences

Now, in Figure 71, we will further examine the covariance differences highlighted in the plot of the second against the fourth and fifth SMVCIR coordinates. In the standardized coefficients for the second SMVCIR coordinates, we see high mass on **Alcohol**, **Color**, and **Proline**. In the fourth dimension we see high mass placed on **NFP**. The fifth dimension places high mass on **Mg**, **NFP**, and **Hue**.

We see a covariance difference between **Color** and **Hue** as we move from the Barbera cultivar to the Grignolino or Barolo. The covariance of **Proline** and **Hue** differs in Barolo from Grignolino and Barbera. **Hue** also differs in covariance with **NFP** as we change cultivar groups from Grignolino to Barolo and Barbera. Besides covariance differences with regard to **Hue**, we see a difference in all cultivars in the covariance of **Mg** and **NFP**.

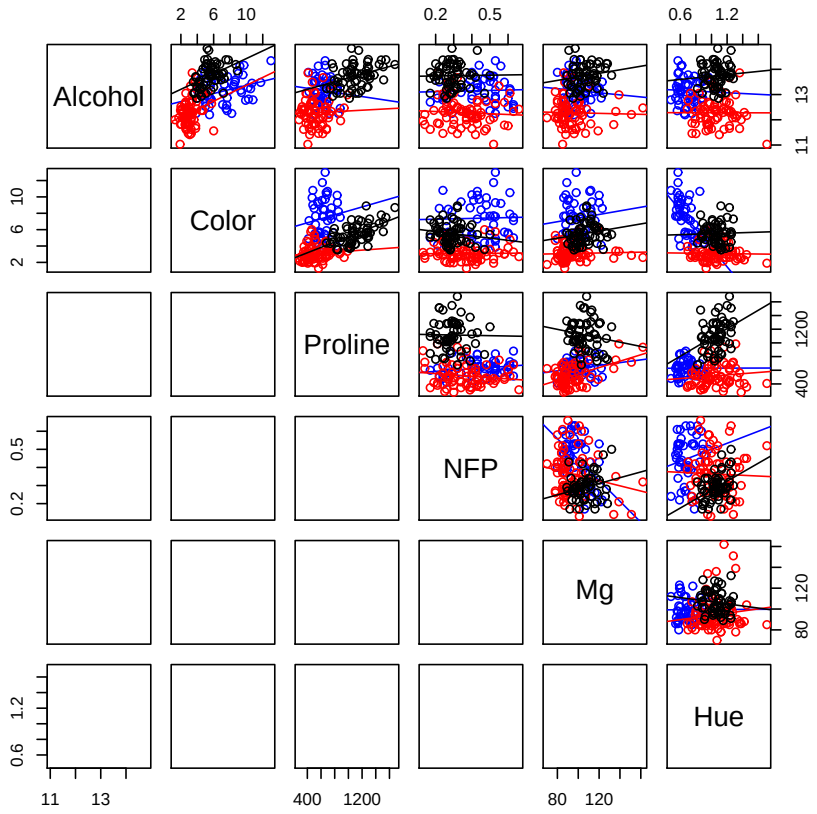

Figure 71. Wine, SMVCIR D_2, D_4, D_5 Covariance Differences

We have found some interesting cultivar group differences by using SMVCIR on the wine data. Now we will see how SAVE and SIR perform. SIR finds two location differences between the cultivars in Figure 72. It separates the groups better than SMVCIR does, but it will provide less covariance difference information. Using the iterated test scheme (4.1), SAVE finds only 7 differences. We examine the first five in Figure 73, and there is little location difference in the three groups, unlike SMVCIR.

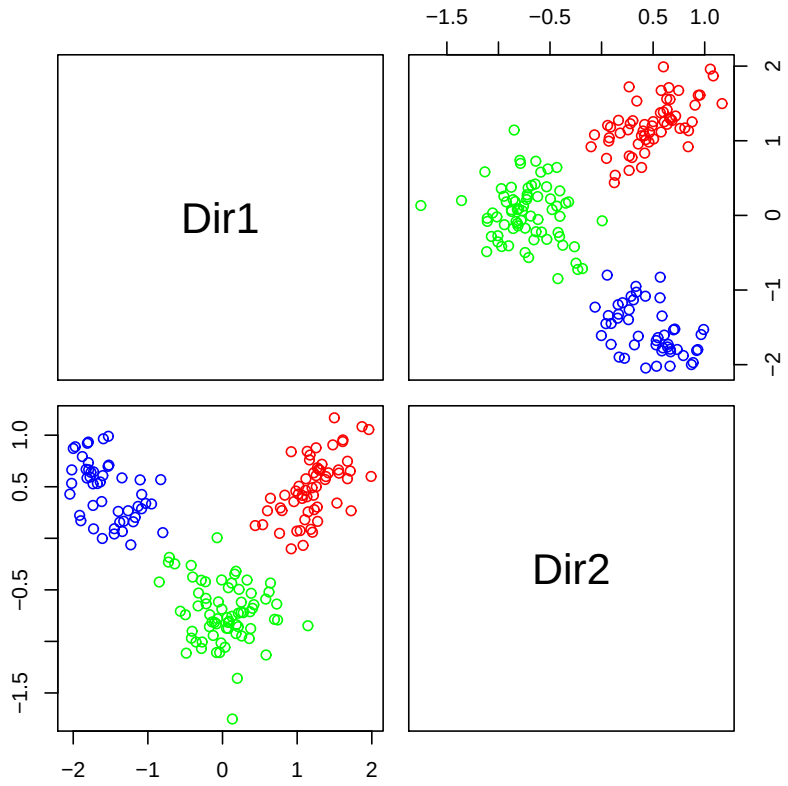

Figure 72. Wine, SIR Coordinates

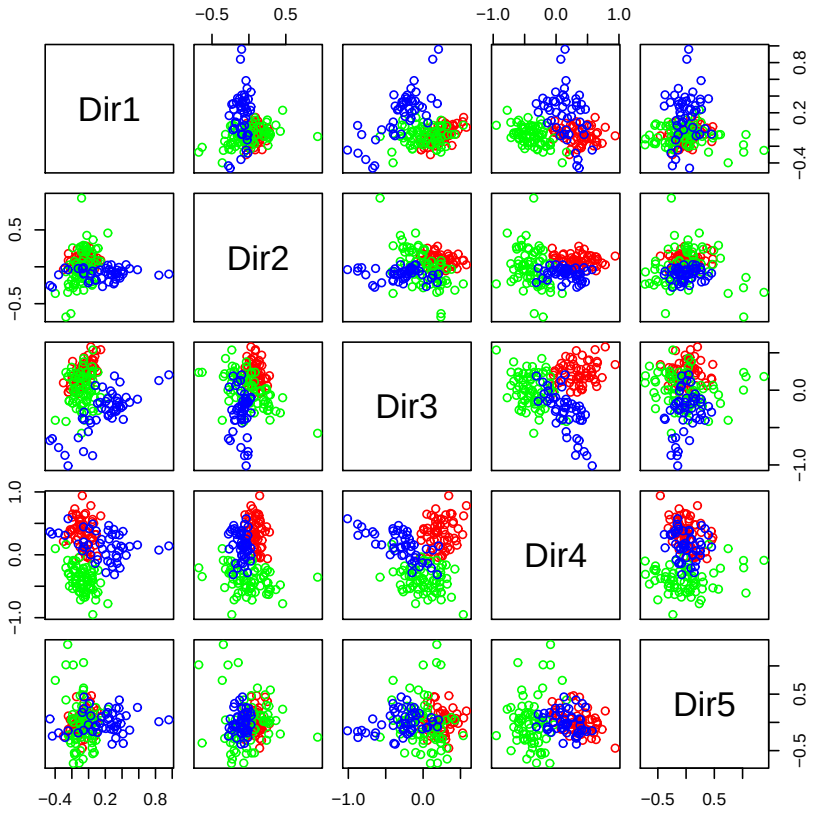

Figure 73. Wine, SAVE Coordinates

CHAPTER VIII

DIRECTIONS OF STUDY

Potential directions for further study of SMVCIR include study of the iterated testing algorithm SMVCIR method (4.1) as a multiple hypothesis testing situation, variable selection for SMVCIR, the application of SMVCIR to continuous response data, and the interaction of the original SMVCIR ordering procedure and analysis method employed in Sheather et al. (2008) with the dimensionality test.

A. Multiple Hypothesis Testing

The first direction we discuss is evaluation of the iterated testing scheme (4.1) as a multiple hypothesis test situation. Suppose we use the iterated testing scheme (4.1) with a certain set test size α for each individual test. Our probability of making a type 1 error when choosing d will not necessarily be α. Suppose we ran k independent tests, then the probability of making a type 1 error would be $k\alpha$. This is not the case here, things are more complicated. We won't even necessarily perform all k tests. We only run test i if we reject the null hypothesis of test $i-1$. This discrepancy between the individual and total test size can cause problems in inference if we do not control for it.

The simulations in Chapter VI showed that SMVCIR did not suffer from this problem. In those situations, the previous tests had great power under the true alternative. In each of the studied situations, the differences were all practically significant. Looking at slight differences, as SAVE might have been doing, we may find that SMVCIR becomes much more sensitive to the individual vs. total test size discrepancy.

Future studies should run simulations of the iterated testing scheme (4.1) with

slight differences. It may also be worthwhile to take a hard theoretic look at the iterated testing scheme (4.1), to see if there formal relationships between the individual and total test sizes that can be used.

B. Variable Selection

As we increase the number of variables, we increase the accuracy of our group discrimination. But at the same time, we decrease the precision of our estimation. Linear regression and other regression modeling techniques face this issue as well. In those situations information criteria are used to evaluate the trade off between increased accuracy and increased model complexity (leading to lower precision).

We might create information criteria for use in SMVCIR. It might also be worthwhile to develop variable selection tests. These tests exist in SAVE and SIR, as shown in Shao et al. (2007). Variable selection tests evaluate the null hypothesis that a variable does not influence the model. For SMVCIR, we could test whether the coefficients of the spanning set eigenvectors corresponding to a particular variable are zero.

Developing variable selection tests and information criteria for SMVCIR are worthwhile endeavors. This development will be performed in later research.

C. Continuous Response

We have restricted our attention to cases where our response variable was either discrete or had fixed slice points before any sampling is performed. This allows us to code the response variable values into groups before any sampling as well.

When we have a continuous response variable, this restriction will cause us to be unable to analyze certain models when we do not have adequate information about

the response. If we keep with this restriction, we could only arbitrarily choose slice points before receiving data and then apply a bad model.

A better solution is to ease the restriction and allow the slice points to be dynamically determined by the data as sampling is performed. We allow an extra level of randomness here, determined by the slicing algorithm we use. We will see that this method can be very useful in the following examples. We see that the SMVCIR output under this dynamic slicing can be very useful. Work remains to be done in applying statistical inference to it.

1. Slice Determination

There are several different ways to perform slicing in the presence of a continuous response. In the following examples we use 10 slices for SMVCIR, creating them so that each slice contains an equal proportion of the data points. This is accomplished by sorting the data points based on the response, and then putting the slice points after a certain number of counted out points.

The dr package (Weisberg (2009)), which we use to calculate SIR and SAVE, takes a user specified "nslices" argument or determines the total number of slices based on the number of predictors. It then performs a similar algorithm to ours, though ties are always placed in the same slice. Our slicing algorithm would be smarter to handle ties in this way, and when we implement it again in later study we will do this. We allow the dr package to automatically determine the number of slices based on the predictors.

2. Examples

We consider two examples. In the first we have a quadratic response y, and five predictors $x_1, \ldots, x_5$.

$$\begin{aligned} x_1, \ldots, x_5 &\sim N(0, 1) \\ \epsilon &\sim N(0, .01) \\ y &= (x_1 + x_2 + x_3 + x_4 + x_5 + \epsilon)^2 \end{aligned} \tag{8.1}$$

We draw 500 samples and allocate 10 slices at equidistant points in the sample distribution of y. After invoking SMVCIR on the sample, we obtain the following first dimension results. By discriminating on the variance and covariance differences of the predictors across the values of y, we uncover the functional form of y given the predictors (Figure 74).

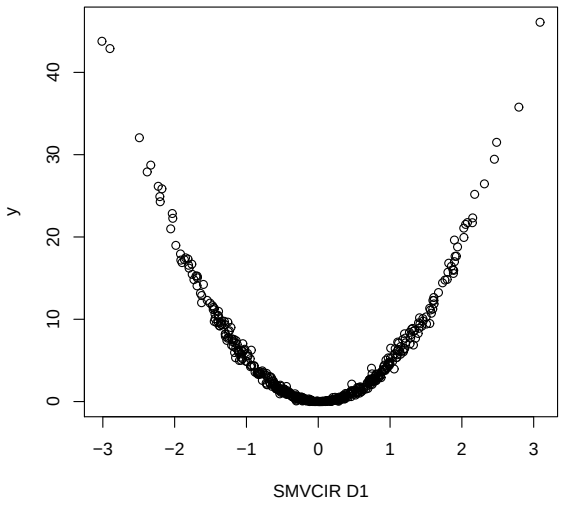

Figure 74. SMVCIR, Quadratic Continuous Example

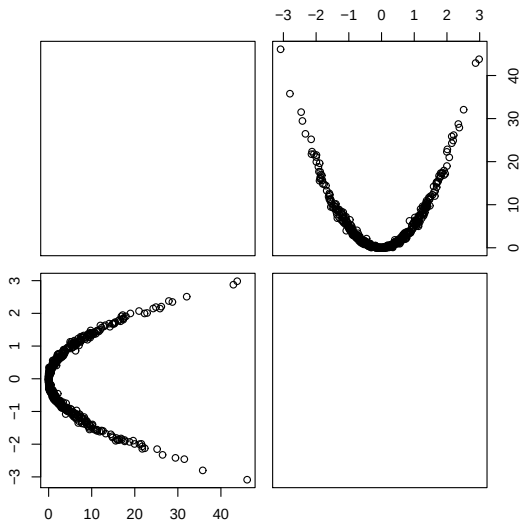

Figure 75. SAVE, Quadratic Continuous Example

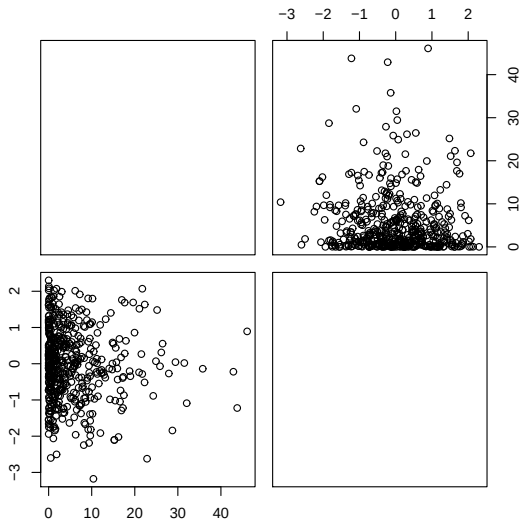

Figure 76. SIR, Quadratic Continuous Example

SAVE is able to find this form as well in Figure 75. We use SAVE's default slicing procedure and the default plotting function from the dr package. We also use the dr package again to see how SIR interprets the situation. Not seeing any mean difference (there should be none), SIR does not do well (Figure 76).

In Figure 77, plots of the original sample predictors and response show little evidence of this form. So the inverse regression algorithms really helped our analysis.

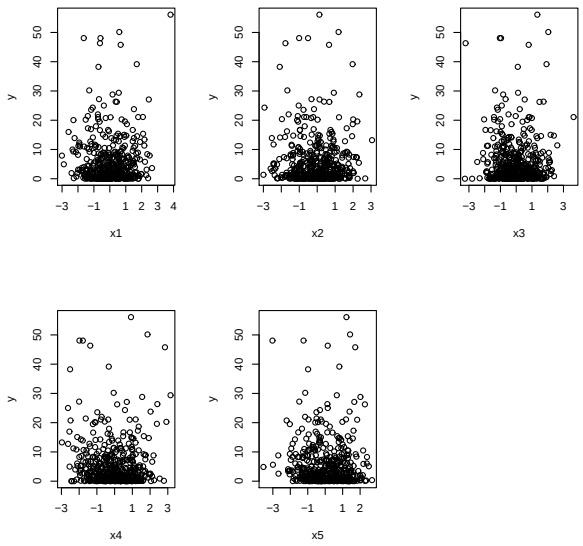

Figure 77. Original Data, Quadratic Continuous Example

Our next example is more complicated. We will use a similar distribution of the predictors and error term from (8.1), but the response is very different.

$$\begin{aligned} &x_1, \ldots, x_{10} \sim N(0,1) \\ &\epsilon \sim N(0, .4^2) \\ &y = \sin(x_1 + \ldots + x_{10}) + \epsilon \\ &\text{rotated counterclockwise of } x_1 + \ldots + x_{10} \text{ axis by an angle of 3} \end{aligned} \tag{8.2}$$

We draw a sample of 1000 points from this population situation. In Figure 78, there is a plot of the rotated y and rotated $x_1 + \ldots + x_{10}$ on the left. On the right is the original y and $x_1 + \ldots + x_{10}$. We input the rotated y and the predictors $x_1, \ldots, x_{10}$ to SMVCIR, SAVE, and, SIR.

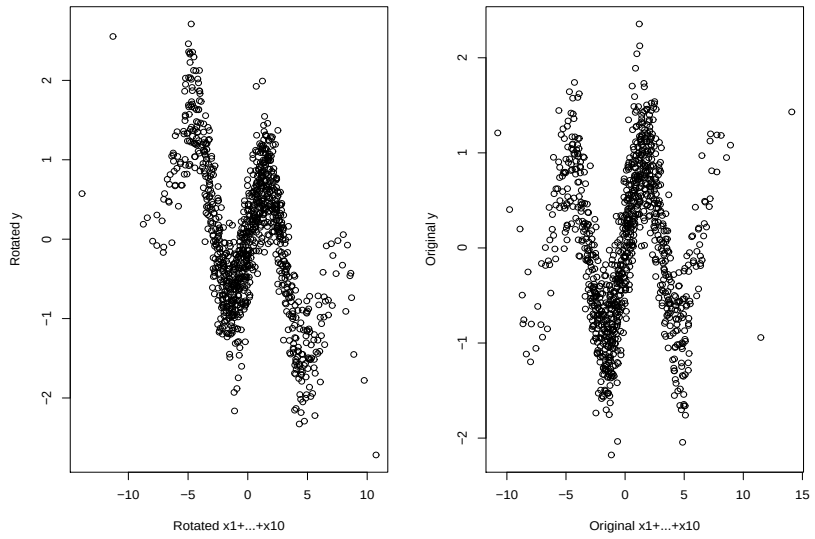

Figure 78. Rotation Example, Rotated And Original Data

Our favorable results under SMVCIR are seen Figure 79. The SAVE results are seen in Figure 80 and the SIR results are given in Figure 81.

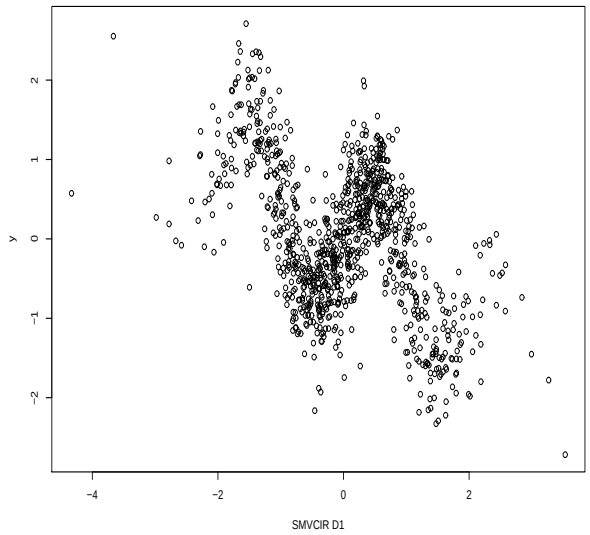

Figure 79. SMVCIR, Diagonal Sine Example

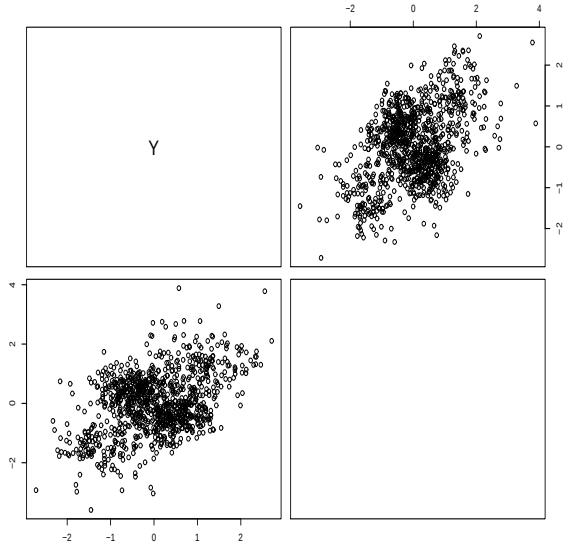

Figure 80. SAVE, Diagonal Sine Example

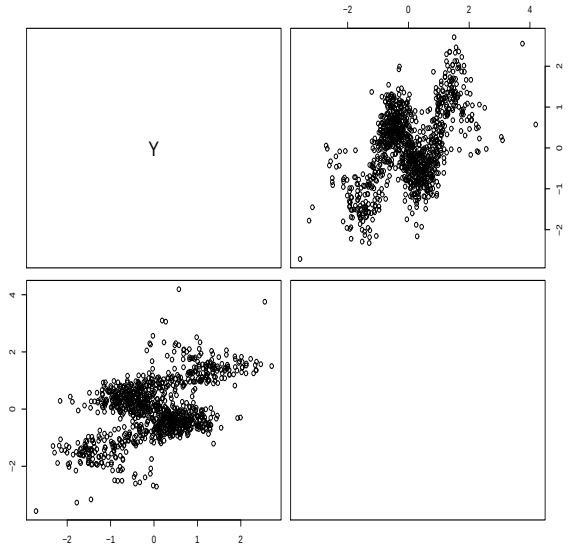

Figure 81. SIR, Diagonal Sine Example

We see that SIR and SMVCIR properly find the shape relating the predictors $x_1, \ldots, x_{10}$ and y. SAVE does not do very well. A matrix plot of the original sample predictors and response in Figure 82 shows little evidence of diagonal sine form relationship between them. So the inverse regression algorithms really helped our analysis again.

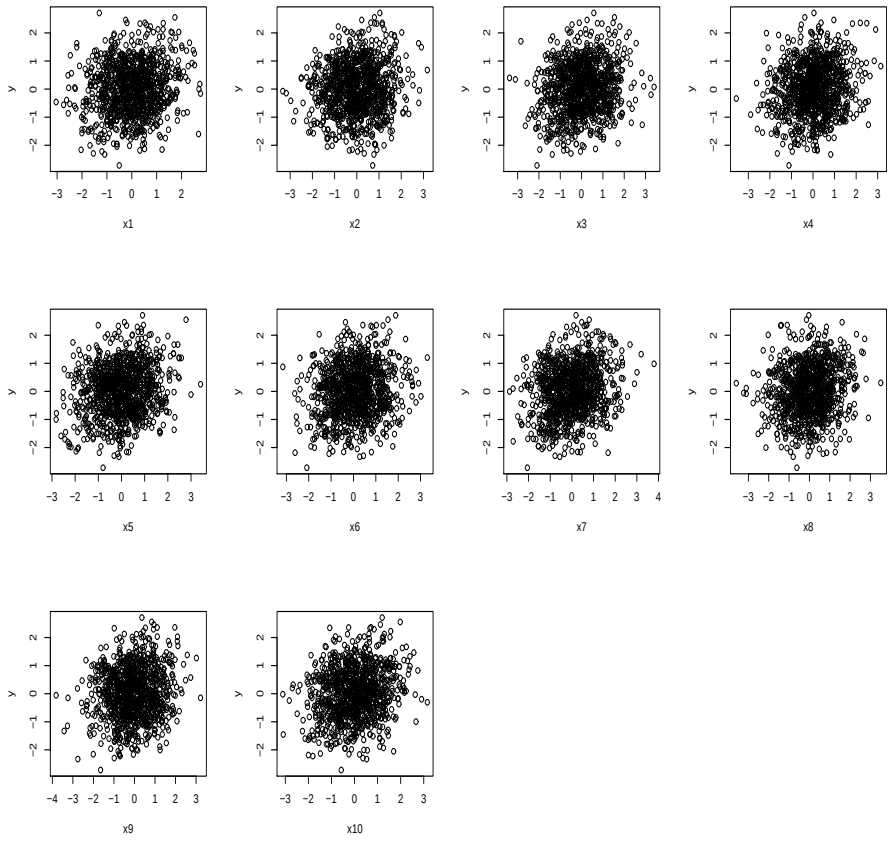

Figure 82. Original Data, Diagonal Sine Example

These two examples have demonstrated how helpful SMVCIR can be in the context of a continuous response. There is clear motivation to develop the dimensionality test to handle dynamically determined slice points and group categories.

D. Original SMVCIR Ordering Procedure

The original SMVCIR ordering procedure and analysis method (from Sheather et al. (2008)) was briefly described in Chapter I. Here we approximated the SMVCIR kernel using only a subset of the SMVCIR spanning set vectors. We order the spanning set vectors by their relative linear independence and then form a restricted spanning set using the first $r \leq k$ most independent vectors. The quantity r was chosen using a scree plot method, where r singular values of the SMVCIR spanning set produce some percentage of its total discriminating power (measured as a sum of the singular values).

The kernel is then formed as cross product of the restricted spanning set rather than the full spanning set. It follows that the rank of the kernel and thus the number of SMVCIR dimensions is less than or equal to r. So examination is restricted to the first r dimensions.

The method of choosing r incorporates the practical significance of the SMVCIR spanning set vectors. In the examples of Chapter VII, we used a similar tactic of selecting the number of dimensions for viewing by examining a scree plot of the singular values. We found in Chapter VII that the dimension chosen by our dimensionality test using 4.1 served as a useful upper bound for the practically significant dimension (how many plots we wished to examine). Perhaps it could serve as a useful upper bound for r the number of spanning set vectors to be used in forming the kernel under the original SMVCIR method (Sheather et al. (2008)). The original SMVCIR method was also only employed using the Mahalanobis standardization rather than the scaling standardization. It is worthwhile to investigate how it fares under the scaling standardization.

CHAPTER IX

SUMMARY

We have provided an augmented SMVCIR algorithm, providing a new standardization method and a dimensionality test. Through examples and simulations we have shown how the new SMVCIR algorithm works and provides useful output for statistical modeling.

We began by providing a detailed explanation of the original SMVCIR algorithm. This included an introduction to the SAVE and SIR inverse regression algorithms.

Then we discussed the new scaling standardization. It was demonstrated how the Mahalanobis standardization could add extra group differences in variance and covariance. These extra differences would add noise dimensions to the SMVCIR space and make any dimensionality test difficult to interpret. We showed how the scaling standardization would not add extra variance or covariance differences.

After we established the appropriateness of the scaling standardization, we developed the dimensionality test for the SMVCIR space. We spent a chapter formulating the distribution of the SMVCIR spanning set as asymptotically multivariate normal (3.82). Then we used this distribution and asymptotic properties of the eigen and singular value decompositions to form our final dimensionality test statistic (4.2) and its null distribution. We found three asymptotically equivalent null distributions for the final test statistic: theory (4.27), empirical (4.28), and approximate empirical (4.30). Use of each distribution results in a slightly different test, though they are asymptotically equivalent. This test would be applied in an iterated scheme (4.1), starting at $d = 0$ and testing one dimension after the other until a null hypothesis dimension is accepted. This is the estimated dimension of the SMVCIR space.

Next we provided an example of how SMVCIR and the dimensionality test are

performed. We also provided extra computational details and outlined how we would evaluate the performance of the SMVCIR dimensionality test. We described some of the techniques in Chris J. Lloyd's paper Lloyd (2005) and discussed how we would apply them to our evaluations.

We then performed vigorous simulations of the SMVCIR dimensionality test statistic. Three general multivariate populations were used: normal, Student's T, and exponential. We applied a variety of mean, variance, and covariance difference situations to these populations and evaluated the SMVCIR dimensionality tests in each context. The empirical (4.28) and approximate empirical (4.30) tests were studied in each population setting first. The test size was evaluated at small and large sample sizes. Satisfactory results were obtained. The tests' power in detecting small differences of each type (mean, variance, and covariance) was also studied. Lloyd's techniques were employed to compare the performance of the approximate empirical (4.30) and empirical (4.28) with themselves and each other. Generally we found that the empirical and approximate empirical were approximately equivalent. They were both relatively powerful as well. The theory (4.27) null reference distribution was studied as well, using multivariate normal and mixture multivariate normal populations. In the studied settings, we found that the actual distribution of the SMVCIR test statistic did converge to the theory distribution once we let the sample size become suitably large.

In our simulation study we also performed informal simulations to evaluate the iterated testing scheme (4.1). We used the same population and mean, variance, and covariance situations as we used in evaluating the SMVCIR test sizes. In these simulations both the empirical (4.28) and the approximate empirical (4.30) SMVCIR tests were used. SAVE and SIR were computed on each simulated sample as well. We found that both SMVCIR tests and SIR performed as they should. SAVE had some

difficulties choosing a single dimension with regularity in some of the situations.

Once we completed our simulations, we used SMVCIR on some real world examples. We begin with the pen digit data from Zhu and Hastie (2003). Then we used several datasets from (Izenman (2008)). SAVE and SIR were used on every studied example as well. We found SMVCIR performed well and was competitive with SAVE and SIR.

Finally, we outlined potential directions for further research. Study of the iterated testing scheme (4.1) as a multiple hypothesis situation was one direction. Variable selection on the SMVCIR predictors was another useful subject for further study. we took a preliminary look at the use of SMVCIR on continuous response population settings. We compared how SMVCIR performed against SAVE and SIR in uncovering the functional form relating the response and predictors in two situations. In the first situation, the response was a quadratic function of the summed predictors. Here SMVCIR performed as well as SAVE, and both outperformed SAVE. In the second situation, the response was a trigonometric function (sin) of the summed predictors. Adding further complexity, we rotated the response so that the sine wave was diagonal rather than horizontal. In this situation SMVCIR performed as well as SIR, and both outperformed SAVE. We finally briefly examined how we might research the interaction of the SMVCIR dimensionality test with the original SMVCIR ordering procedure and dimension choice method.

There is more research to be done on SMVCIR. There is also much more practical application of SMVCIR to be performed. The R programs used to perform the examples and simulations in this dissertation will be used to make and redistributable R package. By using this package, individuals can perform SMVCIR on their own datasets and obtain useful results. Conversion of the R programs to other languages, such as SAS and Stata may also be performed.

REFERENCES

Adler, D., and Murdoch, D. (2009), *rgl: 3D Visualization Device System (OpenGL).* R Package Version 0.84.
URL: *http://CRAN.R-project.org/package=rgl*

Bentler, P. M., and Xie, J. (2000), "Corrections to Test Statistics in Principal Hessian Directions," *Statistics and Probability Letters*, 47, 381–389.

Bura, E., and Pfeiffer, R. (2008), "On the Distribution of the Left Singular Vectors of a Random Matrix and Its Applications," *Statistics and Probability Letters*, 78, 2275–2280.

Clopper, C. J., and Pearson, E. S. (1935), "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," *Biometrika*, 26, 404–413.

Cook, R. D. (1998), *Regression Graphics: Ideas for Studying Regressions Through Graphics*, New York: Wiley.

Cook, R. D. (2000), "SAVE: A Method for Dimension Reduction and Graphics in Regression," *Communications in Statistics - Theory and Methods*, 29, 2109–2121.

Cook, R. D., and Yin, X. (2001), "Dimension Reduction and Visualization in Discriminant Analysis," *Aust. N. Z. J. Stat*, 43, 147–199.

Eaton, M. L. (1983), *Multivariate Statistics - A Vector Space Approach*, New York: Wiley.

Eaton, M. L., and Tyler, D. E. (1994), "Asymptotic Distributions of Singular Val-

ues with Applications to Canonical Correlations and Correspondence Analysis," *Journal of Multivariate Analysis*, 50, 238–264.

Gannoun, A., and Saracco, J. (2003), "An Asymptotic Theory for SIR_α Method," *Statistica Sinica*, 13, 297–310.

Golub, G. H., and Loan, C. F. V. (1996), *Matrix Computations* (3rd ed.), Baltimore, Maryland: The John Hopkins University Press.

Guttman, I. (1982), *Linear Models: An Introduction*, New York: Wiley.

Izenman, A. J. (2008), *Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning*, New York: Springer.

Li, K. C. (1991), "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association*, 86, 316–327.

Lloyd, C. J. (2005), "Estimating Test Power Adjusted for Size," *Journal of Statistical Computation and Simulation*, 75, 921–933.

Lütkepohl, H. (2007), *New Introduction to Multiple Time Series Analysis* (2nd ed.), New York: Springer.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), *Multivariate Analysis*, New York: Academic Press.

Marron, J. S., and Wand, M. P. (1992), "Exact Mean Integrated Squared Error," *Annals of Statistics*, 20, 712–736.

Muirhead, R. J. (1982), *Aspects of Multivariate Statistical Analysis*, New York: Wiley.

R Development Core Team (2009), *R: A Language and Environment for Statistical Computing*, Vienna, Austria, R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, New York: Wiley.

Shao, Y., Cook, R. D., and Weisberg, S. (2007), "Marginal Tests with Sliced Average Variance Estimation," *Biometrika*, 94, 285–296.

Sheather, S. J., McKean, J. W., and Crimin, K. (2008), "Sliced Mean Variance-Covariance Inverse Regression," *Computation Statistics & Data Analysis*, 52, 1908–1927.

Tyler, D. E. (1981), "Asymptotic Inference for Eigenvectors," *Annals of Statistics*, 9, 725–736.

Venables, W. N., and Ripley, B. D. (2002), *Modern Applied Statistics with S* (4th ed.), New York: Springer. ISBN 0-387-95457-0.
URL: *http://www.stats.ox.ac.uk/pub/MASS4*

Weisberg, S. (2009), *dr: Methods for Dimension Reduction for Regression*. R Package Version 3.0.4.
URL: *http://CRAN.R-project.org/package=dr*

Yin, X. (2005), Tests of Dimension with Sliced Average Variance Estimation,, Research Report 13, Department of Statistics, University of Georgia.

Zhu, M., and Hastie, T. J. (2003), "Feature Extraction for Nonparametric Discriminant Analysis," *Journal of Computational and Graphical Statistics*, 12, 101–120.

APPENDIX A

MULTIVARIATE NORMAL PRODUCT MOMENT IDENTITIES

Let $\mathbf{h}$ be a k variate multivariate normal random vector mean $\boldsymbol{\mu}^{\mathbf{h}}$ and covariance matrix $\boldsymbol{\Sigma}^{\mathbf{h}}$. We can partition $\mathbf{h}$ into a $q < k$ variate subvector $\mathbf{h_1}$ and $k-q$ subvector $\mathbf{h_2}$.

$$\begin{aligned} \mathbf{h} &= \begin{bmatrix} \mathbf{h_1} \\ \mathbf{h_2} \end{bmatrix} \\ \boldsymbol{\mu}^{\mathbf{h}} &= \begin{bmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{bmatrix} \\ \boldsymbol{\Sigma}^{\mathbf{h}} &= \begin{bmatrix} \boldsymbol{\Sigma_{11}} & \boldsymbol{\Sigma_{12}} \\ \boldsymbol{\Sigma_{21}} & \boldsymbol{\Sigma_{22}} \end{bmatrix} \end{aligned} \tag{A.1}$$

Conditional on $\mathbf{h_2}$,$\mathbf{h_1}$ is multivariate normal with the following mean and variance.

$$\begin{aligned} \boldsymbol{\mu_{1|2}} &= \mu_1 + \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}^{-1}}\left(\mathbf{h_2} - \mu_2\right) \\ \boldsymbol{\Sigma_{1|2}} &= \boldsymbol{\Sigma_{11}} - \boldsymbol{\Sigma_{12}}\boldsymbol{\Sigma_{22}^{-1}}\boldsymbol{\Sigma_{21}} \end{aligned} \tag{A.2}$$

Marginally,$\mathbf{h_1}$ is multivariate normal with mean $\boldsymbol{\mu_1}$ and variance $\boldsymbol{\Sigma_{11}}$.

Now suppose $\mathbf{w} = (w_1, w_2, w_3, w_4)'$ is a four-variate multivariate normal with

mean and variance parameters.

$$\boldsymbol{\mu}^{\mathbf{w}} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ \mu_4 \end{bmatrix} \qquad \boldsymbol{\Sigma}^{\mathbf{w}} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{bmatrix} \tag{A.3}$$

In total, we will calculate 10 moment identities. Through generalizing these moments (changing the indices) we will obtain elements all potential product moments that we will need to calculate the theory null distribution (4.27). The first and second order moments are already well known identities, but we include them for completeness.

$$\begin{aligned} &\mathrm{E}\left[w_4\right] \\ &\mathrm{E}\left[w_4^2\right] \\ &\mathrm{E}\left[w_4 w_3\right] \\ &\mathrm{E}\left[w_4^4\right] \\ &\mathrm{E}\left[w_3^3 w_4\right] \\ &\mathrm{E}\left[w_2 w_3 w_4^2\right] \\ &\mathrm{E}\left[w_1 w_2 w_3 w_4\right] \\ &\mathrm{E}\left[w_4^3\right] \\ &\mathrm{E}\left[w_4^2 w_3\right] \\ &\mathrm{E}\left[w_1 w_2 w_3\right] \end{aligned} \tag{A.4}$$

The first moment identity is obvious.

$$\mathrm{E}\left[w_4\right] = \mu_4 \tag{A.5}$$

The second moment identity utilizes the variance identity (5.16). This is also a well known result.

$$\mathrm{E}\left[w_4^2\right] = \sigma_{44} + \mu_4^2 \tag{A.6}$$

The heterogeneous second order product moment identity also uses (5.16). It is another well known result.

$$\mathrm{E}\left[w_4 w_3\right] = \sigma_{43} + \mu_3\mu_4 \tag{A.7}$$

Now we begin the real work with the expectation of the fourth power of w_4. Initially taking advantage of the location-scale family properties of w_4, we obtain the following.

$$\begin{aligned} &\mathrm{E}\left[w_4^4\right] = \\ &\mathrm{E}\left[\left(\sqrt{\sigma_{44}}Z + \mu_4\right)^4\right] \end{aligned} \tag{A.8}$$

We have reduced the problem to taking the expectation of a quartic polynomial of a standard normal variate. We use Pascal's triangle to expand the polynomial. Then we use the linearity of the expectation operator and the fact that odd moments of the standard normal are 0 (since it is symmetric about 0). We will use these tricks repeatedly.

$$\begin{aligned} &\mathrm{E}\left[\sigma_{44}^2 Z^4 + 4\sigma_{44}^{3/2} Z^3\mu_4 + 6\sigma_{44}Z^2\mu_4^2 + 4\sqrt{\sigma_{44}}Z\mu_4^3 + \mu_4^4\right] = \\ &\sigma_{44}^2\mathrm{E}\left[Z^4\right] + 6\sigma_{44}\mu_4^2\mathrm{E}\left[Z^2\right] + \mu_4^4 \end{aligned} \tag{A.9}$$

Now we realize that Z^2 is a $\chi^2_{p=1}$ random variable. So it has mean $p = 1$ and

variance $2p = 2$. So $\mathrm{E}\left[Z^4\right] = \mathrm{Var}\left[Z^4\right] + \mathrm{E}\left[Z^2\right]^2 = 2 + 1 = 3$ and we have the following.

$$\begin{aligned} &\mathrm{E}\left[w_4^4\right] = \\ &\sigma_{44}^2 3 + 6\sigma_{44}\mu_4^2 + \mu_4^4 = \\ &\mu_4^4 + 6\mu_4^2\sigma_{44} + 3\sigma_{44}^2 \end{aligned} \tag{A.10}$$

Now we will calculate the third moment of w_4.

$$\begin{aligned} &\mathrm{E}\left[w_4^3\right] = \\ &\mathrm{E}\left[\left(\sqrt{\sigma_{44}}Z + \mu_4\right)^3\right] = \\ &\mathrm{E}\left[\sigma_{44}^{3/2}Z^3 + 3\sigma_{44}Z^2\mu_4 + 3\sqrt{\sigma_{44}}Z\mu_4^2 + \mu_4^3\right] = \\ &\mathrm{E}\left[3\sigma_{44}Z^2\mu_4 + \mu_4^3\right] = \\ &3\sigma_{44}\mathrm{E}\left[Z^2\right]\mu_4 + \mu_4^3 = \\ &3\sigma_{44}\mu_4 + \mu_4^3 = \\ &\mu_4\left(\mu_4^2 + 3\sigma_{44}\right) \end{aligned} \tag{A.11}$$

Next we calculate the product moment, $\mathrm{E}\left[w_3^3 w_4\right]$.

$$\begin{aligned} &\mathrm{E}\left[w_3^3 w_4\right] = \\ &\mathrm{E}\left[\mathrm{E}\left[w_3^3 w_4 | w_4\right]\right] = \\ &\mathrm{E}\left[w_4\mathrm{E}\left[w_3^3 | w_4\right]\right] \end{aligned} \tag{A.12}$$

We use (A.2) now. Conditioning on w_4, w_3 is normal with mean and variance given below.

$$\begin{aligned} &\mathrm{E}\left[w_3 | w_4\right] = \mu_3 + \tfrac{\sigma_{34}}{\sigma_{44}}\left(w_4 - \mu_4\right) \\ &\mathrm{Var}\left[w_3 | w_4\right] = \sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43} \end{aligned} \tag{A.13}$$

So now we can rewrite the conditional expectation inside of (A.12) in a similar

manner to (A.2).

$$\begin{array}{l} \mathrm{E}\left[w_3^3|w_4\right] = \\ \left(\mu_3 + \frac{\sigma_{34}}{\sigma_{44}}\left(w_4 - \mu_4\right)\right)\left(\begin{array}{l} \left(\mu_3 + \frac{\sigma_{34}}{\sigma_{44}}\left(w_4 - \mu_4\right)\right)^2 + \\ 3\left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43}\right) \end{array}\right) = \\ \left(\mu_3 + \frac{\sigma_{34}}{\sigma_{44}}\left(w_4 - \mu_4\right)\right)^3 + \\ \left(\mu_3 + \frac{\sigma_{34}}{\sigma_{44}}\left(w_4 - \mu_4\right)\right) 3\left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43}\right) \end{array} \tag{A.14}$$

First we will expand the second term above enough to separate w_4 from the rest.

$$\begin{array}{l} \left(\mu_3 + \frac{\sigma_{34}}{\sigma_{44}}\left(w_4 - \mu_4\right)\right) 3\left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43}\right) = \\ 3\mu_3\left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43}\right) \\ -3\left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43}\right)\frac{\sigma_{34}}{\sigma_{44}}\mu_4 \\ +3\left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43}\right)\frac{\sigma_{34}}{\sigma_{44}}w_4 \end{array} \tag{A.15}$$

Now we focus on the cubic term in (A.14).

$$\begin{array}{l} \left(\mu_3 + \frac{\sigma_{34}}{\sigma_{44}}\left(w_4 - \mu_4\right)\right)^3 = \\ \left(\frac{\sigma_{34}}{\sigma_{44}}w_4 + \left(\mu_3 - \frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)\right)^3 = \\ \frac{\sigma_{34}^3}{\sigma_{44}^3}w_4^3 + 3\frac{\sigma_{34}^2}{\sigma_{44}^2}\left(\mu_3 - \frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)w_4^2 \\ +3\frac{\sigma_{34}}{\sigma_{44}}\left(\mu_3 - \frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^2 w_4 \\ +\left(\mu_3 - \frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^3 \end{array} \tag{A.16}$$

Now we return to (A.14) and rewrite it in the expanded form.

$$\begin{aligned}
&\mathrm{E}\left[w_3^3|w_4\right] = \\
&\tfrac{\sigma_{34}^3}{\sigma_{44}^3}w_4^3 + 3\tfrac{\sigma_{34}^2}{\sigma_{44}^2}\left(\mu_3 - \tfrac{\sigma_{34}}{\sigma_{44}}\mu_4\right)w_4^2 \\
&+3\tfrac{\sigma_{34}}{\sigma_{44}}\left(\mu_3 - \tfrac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^2 w_4 \\
&+\left(\mu_3 - \tfrac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^3 \\
&+3\mu_3\left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43}\right) \\
&-3\left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43}\right)\tfrac{\sigma_{34}}{\sigma_{44}}\mu_4 \\
&+3\left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43}\right)\tfrac{\sigma_{34}}{\sigma_{44}}w_4
\end{aligned} \tag{A.17}$$

So now we get the following.

$$\begin{aligned}
&w_4\mathrm{E}\left[w_3^3|w_4\right] = \\
&\tfrac{\sigma_{34}^3}{\sigma_{44}^3}w_4^4 + 3\tfrac{\sigma_{34}^2}{\sigma_{44}^2}\left(\mu_3 - \tfrac{\sigma_{34}}{\sigma_{44}}\mu_4\right)w_4^3 \\
&+3\tfrac{\sigma_{34}}{\sigma_{44}}\left(\mu_3 - \tfrac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^2 w_4^2 \\
&+\left(\mu_3 - \tfrac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^3 w_4 \\
&+3\mu_3\left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43}\right)w_4 \\
&-3\left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43}\right)\tfrac{\sigma_{34}}{\sigma_{44}}\mu_4 w_4 \\
&+3\left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43}\right)\tfrac{\sigma_{34}}{\sigma_{44}}w_4^2
\end{aligned} \tag{A.18}$$

We take the expectation of (A.18) to find (A.12). The third moment identity in

(A.11) and the fourth moment identity in (A.10) are used.

$$
\begin{aligned}
&\mathrm{E}\left[w_4 w_3^3\right] = \\
&\mathrm{E}\left[w_4 \mathrm{E}\left[w_3^3 | w_4\right]\right] = \\
&\tfrac{\sigma_{34}^3}{\sigma_{44}^3}\left(\mu_4^4 + 6\mu_4^2\sigma_{44} + 3\sigma_{44}^2\right) \\
&+3\tfrac{\sigma_{34}^2}{\sigma_{44}^2}\left(\mu_3 - \tfrac{\sigma_{34}}{\sigma_{44}}\mu_4\right)\left(\mu_4\left(\mu_4^2 + 3\sigma_{44}\right)\right) \\
&+3\tfrac{\sigma_{34}}{\sigma_{44}}\left(\mu_3 - \tfrac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^2\left(\sigma_{44} + \mu_4^2\right) \\
&+\left(\mu_3 - \tfrac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^3\mu_4 \\
&+3\mu_3\left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43}\right)\mu_4 \\
&-3\left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43}\right)\tfrac{\sigma_{34}}{\sigma_{44}}\mu_4^2 \\
&+3\left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43}\right)\tfrac{\sigma_{34}}{\sigma_{44}}\left(\sigma_{44} + \mu_4^2\right)
\end{aligned}
\tag{A.19}
$$

We have prepared marginal fourth (A.10) and third (A.11) order moment identities. We have also shown a product moment identity where one of the factors is cubed and the other is linear (A.19).

We have three more fourth order product moment identities to show. The first has two quadratic factors. In our initial derivations we use the conditional moment

identities in (A.13).

$$
\begin{aligned}
&\mathrm{E}\left[w_4^2 w_3^2\right] = \\
&\mathrm{E}\left[\mathrm{E}\left[w_4^2 w_3^2 | w_4\right]\right] = \\
&\mathrm{E}\left[w_4^2 \mathrm{E}\left[w_3^2 | w_4\right]\right] = \\
&\mathrm{E}\left[w_4^2 \left(\mathrm{Var}\left[w_3|w_4\right] + \mathrm{E}\left[w_3|w_4\right]^2\right)\right] = \\
&\mathrm{E}\left[w_4^2 \left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43} + \left(\mu_3 + \tfrac{\sigma_{34}}{\sigma_{44}}\left(w_4 - \mu_4\right)\right)^2\right)\right] = \\
&\mathrm{E}\left[w_4^2 \left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43} + \left(\tfrac{\sigma_{34}}{\sigma_{44}} w_4 + \left(\mu_3 - \tfrac{\sigma_{34}}{\sigma_{44}}\mu_4\right)\right)^2\right)\right] = \\
&\mathrm{E}\left[w_4^2 \left(\begin{array}{l} \sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43} + \frac{\sigma_{34}^2}{\sigma_{44}^2} w_4^2 + \\ 2\left(\mu_3 - \frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)\frac{\sigma_{34}}{\sigma_{44}} w_4 + \left(\mu_3 - \frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^2 \end{array}\right)\right] = \\
&\mathrm{E}\left[w_4^2 \left(\begin{array}{l} \sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43} + \left(\mu_3 - \frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^2 + \\ 2\left(\mu_3 - \frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)\frac{\sigma_{34}}{\sigma_{44}} w_4 + \frac{\sigma_{34}^2}{\sigma_{44}^2} w_4^2 \end{array}\right)\right] =
\end{aligned}
\tag{A.20}
$$

Now we distribute the w_4^2 term and take the expectation. We utilize the (A.10) and (A.11) moment identities.

$$
\begin{aligned}
&\mathrm{E}\left[w_4^2 w_3^2\right] = \\
&\mathrm{E}\left[\begin{array}{l} \left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43} + \left(\mu_3 - \frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^2\right) w_4^2 + \\ 2\left(\mu_3 - \frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)\frac{\sigma_{34}}{\sigma_{44}} w_4^3 + \frac{\sigma_{34}^2}{\sigma_{44}^2} w_4^4 \end{array}\right] = \\
&\mathrm{E}\left[\left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43} + \left(\mu_3 - \tfrac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^2\right) w_4^2\right] \\
&+\mathrm{E}\left[2\left(\mu_3 - \tfrac{\sigma_{34}}{\sigma_{44}}\mu_4\right)\tfrac{\sigma_{34}}{\sigma_{44}} w_4^3\right] \\
&+\mathrm{E}\left[\tfrac{\sigma_{34}^2}{\sigma_{44}^2} w_4^4\right] = \\
&\left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43} + \left(\mu_3 - \tfrac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^2\right)\left(\sigma_{44} + \mu_4^2\right) \\
&+2\left(\mu_3 - \tfrac{\sigma_{34}}{\sigma_{44}}\mu_4\right)\tfrac{\sigma_{34}}{\sigma_{44}}\left(\mu_4\left(\mu_4^2 + 3\sigma_{44}\right)\right) \\
&+\tfrac{\sigma_{34}^2}{\sigma_{44}^2}\left(\mu_4^4 + 6\mu_4^2\sigma_{44} + 3\sigma_{44}^2\right)
\end{aligned}
\tag{A.21}
$$

Now we are going to derive a fourth order product moment identity that involves w_2, w_3, and w_4, $\mathrm{E}\left[w_4^2 w_3 w_2\right]$. This identity involves three variables, so we cannot directly re-use (A.13). We will use the more general (A.2).

$$\begin{aligned} &\mathrm{E}\left[w_4^2 w_3 w_2\right] = \\ &\mathrm{E}\left[\mathrm{E}\left[w_4^2 w_3 w_2 | w_4, w_3\right]\right] = \\ &\mathrm{E}\left[\mathrm{E}\left[\mathrm{E}\left[w_4^2 w_3 w_2 | w_4, w_3\right] | w_4\right]\right] \end{aligned} \tag{A.22}$$

We begin calculations by examining the innermost conditional expectation.

$$\begin{aligned} &\mathrm{E}\left[w_4^2 w_3 w_2 | w_4, w_3\right] = \\ &w_4^2 w_3 \mathrm{E}\left[w_2 | w_4, w_3\right] = \\ &w_4^2 w_3 \left(\mu_2 + \begin{bmatrix} \sigma_{23} & \sigma_{24} \end{bmatrix} \begin{bmatrix} \sigma_{33} & \sigma_{34} \\ \sigma_{43} & \sigma_{44} \end{bmatrix}^{-1} \left(\begin{bmatrix} w_3 \\ w_4 \end{bmatrix} - \begin{bmatrix} \mu_3 \\ \mu_4 \end{bmatrix} \right) \right) = \\ &w_4^2 w_3 \mu_2 + \\ &\begin{bmatrix} \sigma_{23} & \sigma_{24} \end{bmatrix} \begin{bmatrix} \sigma_{33} & \sigma_{34} \\ \sigma_{43} & \sigma_{44} \end{bmatrix}^{-1} \begin{bmatrix} w_4^2 w_3^2 - w_4^2 w_3 \mu_3 \\ w_4^3 w_3 - w_4^2 w_3 \mu_4 \end{bmatrix} \end{aligned} \tag{A.23}$$

The inverse of the covariance matrix of w_3 and w_4 is easily calculated as the following.

$$\begin{bmatrix} \sigma_{33} & \sigma_{34} \\ \sigma_{43} & \sigma_{44} \end{bmatrix}^{-1} = \frac{1}{\sigma_{33}\sigma_{44} - \sigma_{34}\sigma_{43}} \begin{bmatrix} \sigma_{44} & -\sigma_{34} \\ -\sigma_{43} & \sigma_{33} \end{bmatrix} \tag{A.24}$$

Now we plug this back into (A.23).

$$
\begin{aligned}
&\mathrm{E}\left[w_4^2 w_3 w_2 | w_4, w_3\right] = \\
&w_4^2 w_3 \mu_2 + \\
&\frac{1}{\sigma_{33}\sigma_{44} - \sigma_{34}\sigma_{43}} \begin{bmatrix} \sigma_{23} & \sigma_{24} \end{bmatrix} \begin{bmatrix} \sigma_{44} & -\sigma_{34} \\ -\sigma_{43} & \sigma_{33} \end{bmatrix} \begin{bmatrix} w_4^2 w_3^2 - w_4^2 w_3 \mu_3 \\ w_4^3 w_3 - w_4^2 w_3 \mu_4 \end{bmatrix} = \\
&w_4^2 w_3 \mu_2 + \\
&\frac{1}{\sigma_{33}\sigma_{44} - \sigma_{34}\sigma_{43}} \begin{bmatrix} \sigma_{23} & \sigma_{24} \end{bmatrix} \begin{bmatrix} \begin{bmatrix} \sigma_{44} & -\sigma_{34} \end{bmatrix} \begin{bmatrix} w_4^2 w_3^2 - w_4^2 w_3 \mu_3 \\ w_4^3 w_3 - w_4^2 w_3 \mu_4 \end{bmatrix} \\ \begin{bmatrix} -\sigma_{43} & \sigma_{33} \end{bmatrix} \begin{bmatrix} w_4^2 w_3^2 - w_4^2 w_3 \mu_3 \\ w_4^3 w_3 - w_4^2 w_3 \mu_4 \end{bmatrix} \end{bmatrix} = \\
&w_4^2 w_3 \mu_2 + \\
&\frac{1}{\sigma_{33}\sigma_{44} - \sigma_{34}\sigma_{43}} \begin{pmatrix} \begin{bmatrix} \sigma_{23} & \sigma_{24} \end{bmatrix} \\ \begin{bmatrix} \sigma_{44}\left(w_4^2 w_3^2 - w_4^2 w_3 \mu_3\right) - \sigma_{34}\left(w_4^3 w_3 - w_4^2 w_3 \mu_4\right) \\ -\sigma_{43}\left(w_4^2 w_3^2 - w_4^2 w_3 \mu_3\right) + \sigma_{33}\left(w_4^3 w_3 - w_4^2 w_3 \mu_4\right) \end{bmatrix} \end{pmatrix}
\end{aligned}
\tag{A.25}
$$

We complete the final matrix multiplication to obtain the scalar value of the conditional expectation.

$$
\begin{aligned}
&\mathrm{E}\left[w_4^2 w_3 w_2 | w_4, w_3\right] = \\
&w_4^2 w_3 \mu_2 + \\
&\frac{1}{\sigma_{33}\sigma_{44} - \sigma_{34}\sigma_{43}} \begin{pmatrix} \sigma_{23}\sigma_{44}\left(w_4^2 w_3^2 - w_4^2 w_3 \mu_3\right) \\ -\sigma_{23}\sigma_{34}\left(w_4^3 w_3 - w_4^2 w_3 \mu_4\right) \\ -\sigma_{24}\sigma_{43}\left(w_4^2 w_3^2 - w_4^2 w_3 \mu_3\right) \\ +\sigma_{24}\sigma_{33}\left(w_4^3 w_3 - w_4^2 w_3 \mu_4\right) \end{pmatrix}
\end{aligned}
\tag{A.26}
$$

Now we will calculate $\mathrm{E}\left[\mathrm{E}\left[w_4^2 w_3 w_2 | w_4, w_3\right] | w_4\right]$. To simplify this next calculation,

we rearrange (A.26) so that each power of w_3 is clearly identified.

$$\begin{aligned}
&\mathrm{E}\left[w_4^2 w_3 w_2 | w_4, w_3\right] = \\
&\left(\begin{array}{l} w_4^2 \mu_2 \\ \\ \\ \\ \\ \\ \end{array} \right. \\
&\quad \left. + \frac{1}{\sigma_{33}\sigma_{44} - \sigma_{34}\sigma_{43}} \left(\begin{array}{l} -\sigma_{23}\sigma_{44} w_4^2 \mu_3 \\ +\sigma_{23}\sigma_{34} w_4^2 \mu_4 \\ -\sigma_{23}\sigma_{34} w_4^3 \\ +\sigma_{24}\sigma_{43} w_4^2 \mu_3 \\ +\sigma_{24}\sigma_{33} w_4^3 \\ -\sigma_{24}\sigma_{33} w_4^2 \mu_4 \end{array} \right) w_3 \right) \\
&+ \frac{1}{\sigma_{33}\sigma_{44} - \sigma_{34}\sigma_{43}} \left(\begin{array}{l} \sigma_{23}\sigma_{44} w_4^2 \\ -\sigma_{24}\sigma_{43} w_4^2 \end{array} \right) w_3^2
\end{aligned} \tag{A.27}$$

Now we use (A.27) and (A.13) to calculate $\mathrm{E}\left[\mathrm{E}\left[w_4^2 w_3 w_2 | w_4, w_3\right] | w_4\right]$.

$$\begin{aligned}
&\mathrm{E}\left[\mathrm{E}\left[w_4^2 w_3 w_2 | w_4, w_3\right] | w_4\right] = \\
&\left(\begin{array}{l} w_4^2 \mu_2 \\ \\ \\ \\ \\ \\ \end{array} \right. \\
&\quad \left. + \frac{1}{\sigma_{33}\sigma_{44} - \sigma_{34}\sigma_{43}} \left(\begin{array}{l} -\sigma_{23}\sigma_{44} w_4^2 \mu_3 \\ +\sigma_{23}\sigma_{34} w_4^2 \mu_4 \\ -\sigma_{23}\sigma_{34} w_4^3 \\ +\sigma_{24}\sigma_{43} w_4^2 \mu_3 \\ +\sigma_{24}\sigma_{33} w_4^3 \\ -\sigma_{24}\sigma_{33} w_4^2 \mu_4 \end{array} \right) \left(\mu_3 + \tfrac{\sigma_{34}}{\sigma_{44}}\left(w_4 - \mu_4\right)\right) \right) \\
&+ \frac{1}{\sigma_{33}\sigma_{44} - \sigma_{34}\sigma_{43}} \left(\begin{array}{l} \sigma_{23}\sigma_{44} w_4^2 \\ -\sigma_{24}\sigma_{43} w_4^2 \end{array} \right) \\
&\left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43} + \left(\mu_3 + \tfrac{\sigma_{34}}{\sigma_{44}}\left(w_4 - \mu_4\right)\right)^2\right)
\end{aligned} \tag{A.28}$$

We have calculated all the conditional moments to be used in finding $\mathrm{E}\left[w_4^2 w_3 w_2\right]$.

To finalize calculations, we will take the expectation of (A.28). First we expand (A.28) so that the $\mathrm{E}\left[w_3|w_4\right]$ term has been multiplied through.

$$
\begin{aligned}
&\mathrm{E}\left[\mathrm{E}\left[w_4^2 w_3 w_2 | w_4, w_3\right] | w_4\right] = \\
&\left(\begin{array}{l} w_4^2\mu_2 \\ \\ \\ \\ \\ \\ \end{array} \right. \\
&\left. \begin{array}{l} \\ +\frac{1}{\sigma_{33}\sigma_{44}-\sigma_{34}\sigma_{43}} \end{array} \left(\begin{array}{l} -\sigma_{23}\sigma_{44}w_4^2\mu_3 \\ +\sigma_{23}\sigma_{34}w_4^2\mu_4 \\ -\sigma_{23}\sigma_{34}w_4^3 \\ +\sigma_{24}\sigma_{43}w_4^2\mu_3 \\ +\sigma_{24}\sigma_{33}w_4^3 \\ -\sigma_{24}\sigma_{33}w_4^2\mu_4 \end{array} \right) \right) \left(\mu_3 - \tfrac{\sigma_{34}}{\sigma_{44}}\mu_4\right) \\
&+\tfrac{\sigma_{34}}{\sigma_{44}} \left(\begin{array}{l} w_4^3\mu_2 \\ \\ \\ \\ \\ \\ \end{array} \right. \\
&\left. \begin{array}{l} \\ +\frac{1}{\sigma_{33}\sigma_{44}-\sigma_{34}\sigma_{43}} \end{array} \left(\begin{array}{l} -\sigma_{23}\sigma_{44}w_4^3\mu_3 \\ +\sigma_{23}\sigma_{34}w_4^3\mu_4 \\ -\sigma_{23}\sigma_{34}w_4^4 \\ +\sigma_{24}\sigma_{43}w_4^3\mu_3 \\ +\sigma_{24}\sigma_{33}w_4^4 \\ -\sigma_{24}\sigma_{33}w_4^3\mu_4 \end{array} \right) \right) \\
&+\tfrac{1}{\sigma_{33}\sigma_{44}-\sigma_{34}\sigma_{43}} \left(\begin{array}{l} \sigma_{23}\sigma_{44}w_4^2 \\ -\sigma_{24}\sigma_{43}w_4^2 \end{array} \right) \\
&\left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43} + \left(\mu_3 + \tfrac{\sigma_{34}}{\sigma_{44}}\left(w_4 - \mu_4\right)\right)^2\right)
\end{aligned}
\tag{A.29}
$$

We can invoke (A.11) and (A.10) directly to find the expectation of the first two

terms in (A.29). We will now focus on the last additive term of (A.29).

$$
\begin{array}{l}
\frac{1}{\sigma_{33}\sigma_{44}-\sigma_{34}\sigma_{43}} \begin{pmatrix} \sigma_{23}\sigma_{44}w_4^2 \\ -\sigma_{24}\sigma_{43}w_4^2 \end{pmatrix} \\
\left(\sigma_{33}-\sigma_{34}\sigma_{44}^{-1}\sigma_{43}+\left(\mu_3+\frac{\sigma_{34}}{\sigma_{44}}\left(w_4-\mu_4\right)\right)^2\right) = \\
\frac{(\sigma_{23}\sigma_{44}-\sigma_{24}\sigma_{43})}{\sigma_{33}\sigma_{44}-\sigma_{34}\sigma_{43}}w_4^2 \\
\begin{pmatrix} \sigma_{33}-\sigma_{34}\sigma_{44}^{-1}\sigma_{43}+\frac{\sigma_{34}^2}{\sigma_{44}^2}w_4^2+ \\ 2w_4\frac{\sigma_{34}}{\sigma_{44}}\left(\mu_3-\frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)+\left(\mu_3-\frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^2 \end{pmatrix} = \\
\frac{\sigma_{23}\sigma_{44}-\sigma_{24}\sigma_{43}}{\sigma_{33}\sigma_{44}-\sigma_{34}\sigma_{43}} \\
\begin{pmatrix} \left(\sigma_{33}-\sigma_{34}\sigma_{44}^{-1}\sigma_{43}\right)w_4^2+\frac{\sigma_{34}^2}{\sigma_{44}^2}w_4^4+ \\ 2w_4^3\frac{\sigma_{34}}{\sigma_{44}}\left(\mu_3-\frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)+\left(\mu_3-\frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^2 w_4^2 \end{pmatrix}
\end{array}
\tag{A.30}
$$

We substitute this back into (A.29) to get the following.

$$
\begin{aligned}
&\mathrm{E}\left[\mathrm{E}\left[w_4^2 w_3 w_2 | w_4, w_3\right] | w_4\right] = \\
&\left(\begin{array}{l} w_4^2 \mu_2 \\ \\ \\ \\ \\ \\ \end{array} \right. \\
&\quad \left. + \frac{1}{\sigma_{33}\sigma_{44} - \sigma_{34}\sigma_{43}} \left(\begin{array}{l} -\sigma_{23}\sigma_{44} w_4^2 \mu_3 \\ +\sigma_{23}\sigma_{34} w_4^2 \mu_4 \\ -\sigma_{23}\sigma_{34} w_4^3 \\ +\sigma_{24}\sigma_{43} w_4^2 \mu_3 \\ +\sigma_{24}\sigma_{33} w_4^3 \\ -\sigma_{24}\sigma_{33} w_4^2 \mu_4 \end{array} \right) \left(\mu_3 - \tfrac{\sigma_{34}}{\sigma_{44}}\mu_4\right) \right) \\
&+ \tfrac{\sigma_{34}}{\sigma_{44}} \left(\begin{array}{l} w_4^3 \mu_2 \\ \\ \\ \\ \\ \\ \end{array} \right. \\
&\quad \left. + \frac{1}{\sigma_{33}\sigma_{44} - \sigma_{34}\sigma_{43}} \left(\begin{array}{l} -\sigma_{23}\sigma_{44} w_4^3 \mu_3 \\ +\sigma_{23}\sigma_{34} w_4^3 \mu_4 \\ -\sigma_{23}\sigma_{34} w_4^4 \\ +\sigma_{24}\sigma_{43} w_4^3 \mu_3 \\ +\sigma_{24}\sigma_{33} w_4^4 \\ -\sigma_{24}\sigma_{33} w_4^3 \mu_4 \end{array} \right) \right) \\
&+ \tfrac{\sigma_{23}\sigma_{44} - \sigma_{24}\sigma_{43}}{\sigma_{33}\sigma_{44} - \sigma_{34}\sigma_{43}} \\
&\left(\begin{array}{l} \left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43}\right) w_4^2 + \frac{\sigma_{34}^2}{\sigma_{44}^2} w_4^4 + \\ 2 w_4^3 \frac{\sigma_{34}}{\sigma_{44}} \left(\mu_3 - \frac{\sigma_{34}}{\sigma_{44}}\mu_4\right) + \left(\mu_3 - \frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^2 w_4^2 \end{array} \right)
\end{aligned}
\tag{A.31}
$$

We use (A.11) and (A.10) directly to find the expectation of (5.39), and thus $\mathrm{E}\left[w_4^2 w_3 w_2\right]$.

We have three more product moment identities to find, $\mathrm{E}\left[w_4 w_3 w_2\right]$, $\mathrm{E}\left[w_4^2 w_3\right]$,and $\mathrm{E}\left[w_4 w_3 w_2 w_1\right]$. We will calculate $\mathrm{E}\left[w_4 w_3 w_2\right]$. The derivation will mirror our calcula-

tion of $\mathrm{E}\left[w_4^2 w_3 w_2\right]$.

$$\begin{aligned} &\mathrm{E}\left[w_4 w_3 w_2\right] = \\ &\mathrm{E}\left[\mathrm{E}\left[w_4 w_3 w_2 | w_4, w_3\right]\right] = \\ &\mathrm{E}\left[\mathrm{E}\left[\mathrm{E}\left[w_4 w_3 w_2 | w_4, w_3\right] | w_4\right]\right] \end{aligned} \tag{A.32}$$

We start by calculating $\mathrm{E}\left[w_4 w_3 w_2 | w_4, w_3\right]$. Note our work in (A.23) and (A.24).

$$\begin{aligned} &\mathrm{E}\left[w_4 w_3 w_2 | w_4, w_3\right] = \\ &w_4 w_3 \mathrm{E}\left[w_2 | w_4, w_3\right] = \\ &w_4^2 w_3 \left(\mu_2 + \begin{bmatrix} \sigma_{23} & \sigma_{24} \end{bmatrix} \begin{bmatrix} \sigma_{33} & \sigma_{34} \\ \sigma_{43} & \sigma_{44} \end{bmatrix}^{-1} \left(\begin{bmatrix} w_3 \\ w_4 \end{bmatrix} - \begin{bmatrix} \mu_3 \\ \mu_4 \end{bmatrix} \right) \right) = \\ &w_4 w_3 \mu_2 + \\ &\begin{bmatrix} \sigma_{23} & \sigma_{24} \end{bmatrix} \begin{bmatrix} \sigma_{33} & \sigma_{34} \\ \sigma_{43} & \sigma_{44} \end{bmatrix}^{-1} \begin{bmatrix} w_4 w_3^2 - w_4 w_3 \mu_3 \\ w_4^2 w_3 - w_4 w_3 \mu_4 \end{bmatrix} = \\ &w_4 w_3 \mu_2 + \\ &\frac{1}{\sigma_{33}\sigma_{44} - \sigma_{34}\sigma_{43}} \begin{bmatrix} \sigma_{23} & \sigma_{24} \end{bmatrix} \begin{bmatrix} \sigma_{44} & -\sigma_{34} \\ -\sigma_{43} & \sigma_{33} \end{bmatrix} \begin{bmatrix} w_4 w_3^2 - w_4 w_3 \mu_3 \\ w_4^2 w_3 - w_4 w_3 \mu_4 \end{bmatrix} \end{aligned} \tag{A.33}$$

Now we perform the matrix multiplications in the last additive term.

$$
\begin{aligned}
&\mathrm{E}\left[w_4 w_3 w_2 | w_4, w_3\right] = \\
&w_4 w_3 \mu_2 + \\
&\frac{1}{\sigma_{33}\sigma_{44}-\sigma_{34}\sigma_{43}} \begin{bmatrix} \sigma_{23} & \sigma_{24} \end{bmatrix} \begin{bmatrix} \begin{bmatrix} \sigma_{44} & -\sigma_{34} \end{bmatrix} \begin{bmatrix} w_4 w_3^2 - w_4 w_3 \mu_3 \\ w_4^2 w_3 - w_4 w_3 \mu_4 \end{bmatrix} \\ \begin{bmatrix} -\sigma_{43} & \sigma_{33} \end{bmatrix} \begin{bmatrix} w_4 w_3^2 - w_4 w_3 \mu_3 \\ w_4^2 w_3 - w_4 w_3 \mu_4 \end{bmatrix} \end{bmatrix} = \\
&w_4^2 w_3 \mu_2 + \\
&\frac{1}{\sigma_{33}\sigma_{44}-\sigma_{34}\sigma_{43}} \begin{pmatrix} \begin{bmatrix} \sigma_{23} & \sigma_{24} \end{bmatrix} \\ \begin{bmatrix} \sigma_{44}\left(w_4 w_3^2 - w_4 w_3 \mu_3\right) - \sigma_{34}\left(w_4^2 w_3 - w_4 w_3 \mu_4\right) \\ -\sigma_{43}\left(w_4 w_3^2 - w_4 w_3 \mu_3\right) + \sigma_{33}\left(w_4^2 w_3 - w_4 w_3 \mu_4\right) \end{bmatrix} \end{pmatrix} = \\
&w_4 w_3 \mu_2 + \\
&\frac{1}{\sigma_{33}\sigma_{44}-\sigma_{34}\sigma_{43}} \begin{pmatrix} \sigma_{23}\sigma_{44}\left(w_4 w_3^2 - w_4 w_3 \mu_3\right) \\ -\sigma_{23}\sigma_{34}\left(w_4^2 w_3 - w_4 w_3 \mu_4\right) \\ -\sigma_{24}\sigma_{43}\left(w_4 w_3^2 - w_4 w_3 \mu_3\right) \\ +\sigma_{24}\sigma_{33}\left(w_4^2 w_3 - w_4 w_3 \mu_4\right) \end{pmatrix}
\end{aligned}
\tag{A.34}
$$

We will next calculate $\mathrm{E}\left[\mathrm{E}\left[w_4 w_3 w_2 | w_4, w_3\right] | w_4\right]$. For this purpose, we arrange

(A.34) so that the multiples of powers of w_3 are clearly identifiable.

$$\begin{aligned}
&\mathrm{E}\left[w_4 w_3 w_2 | w_4, w_3\right] = \\
&\left(\begin{array}{l} w_4\mu_2 \\ \\ +\frac{1}{\sigma_{33}\sigma_{44}-\sigma_{34}\sigma_{43}} \end{array}\right.
\left(\begin{array}{l} -\sigma_{23}\sigma_{44}w_4\mu_3 \\ +\sigma_{23}\sigma_{34}w_4\mu_4 \\ -\sigma_{23}\sigma_{34}w_4^2 \\ +\sigma_{24}\sigma_{43}w_4\mu_3 \\ +\sigma_{24}\sigma_{33}w_4^2 \\ -\sigma_{24}\sigma_{33}w_4\mu_4 \end{array}\right) w_3 \\
&+\frac{1}{\sigma_{33}\sigma_{44}-\sigma_{34}\sigma_{43}} \left(\begin{array}{l} \sigma_{23}\sigma_{44}w_4 \\ -\sigma_{24}\sigma_{43}w_4 \end{array}\right) w_3^2
\end{aligned} \tag{A.35}$$

Now we calculate $\mathrm{E}\left[\mathrm{E}\left[w_4 w_3 w_2 | w_4, w_3\right] | w_4\right]$.

$$\begin{aligned}
&\mathrm{E}\left[\mathrm{E}\left[w_4 w_3 w_2 | w_4, w_3\right] | w_4\right] = \\
&\left(\begin{array}{l} w_4\mu_2 \\ \\ +\frac{1}{\sigma_{33}\sigma_{44}-\sigma_{34}\sigma_{43}} \end{array}\right.
\left(\begin{array}{l} -\sigma_{23}\sigma_{44}w_4\mu_3 \\ +\sigma_{23}\sigma_{34}w_4\mu_4 \\ -\sigma_{23}\sigma_{34}w_4^2 \\ +\sigma_{24}\sigma_{43}w_4\mu_3 \\ +\sigma_{24}\sigma_{33}w_4^2 \\ -\sigma_{24}\sigma_{33}w_4\mu_4 \end{array}\right) \left.\left(\mu_3 + \tfrac{\sigma_{34}}{\sigma_{44}}\left(w_4 - \mu_4\right)\right)\right) \\
&+\frac{1}{\sigma_{33}\sigma_{44}-\sigma_{34}\sigma_{43}} \left(\begin{array}{l} \sigma_{23}\sigma_{44}w_4 \\ -\sigma_{24}\sigma_{43}w_4 \end{array}\right) \\
&\left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43} + \left(\mu_3 + \tfrac{\sigma_{34}}{\sigma_{44}}\left(w_4 - \mu_4\right)\right)^2\right)
\end{aligned} \tag{A.36}$$

Now we multiply through the first additive term.

$$
\begin{aligned}
&\mathrm{E}\left[\mathrm{E}\left[w_4 w_3 w_2 | w_4, w_3\right] | w_4\right] = \\
&\left(\begin{array}{l} w_4\mu_2 \\ \\ \\ \\ \\ \\ \end{array} \right. \\
&\quad \left. + \frac{1}{\sigma_{33}\sigma_{44} - \sigma_{34}\sigma_{43}} \left(\begin{array}{l} -\sigma_{23}\sigma_{44} w_4 \mu_3 \\ +\sigma_{23}\sigma_{34} w_4 \mu_4 \\ -\sigma_{23}\sigma_{34} w_4^2 \\ +\sigma_{24}\sigma_{43} w_4 \mu_3 \\ +\sigma_{24}\sigma_{33} w_4^2 \\ -\sigma_{24}\sigma_{33} w_4 \mu_4 \end{array} \right) \right) \left(\mu_3 - \tfrac{\sigma_{34}}{\sigma_{44}}\mu_4\right) \\
&+ \tfrac{\sigma_{34}}{\sigma_{44}} \left(\begin{array}{l} w_4^2\mu_2 \\ \\ \\ \\ \\ \\ \end{array} \right. \\
&\quad \left. + \frac{1}{\sigma_{33}\sigma_{44} - \sigma_{34}\sigma_{43}} \left(\begin{array}{l} -\sigma_{23}\sigma_{44} w_4^2 \mu_3 \\ +\sigma_{23}\sigma_{34} w_4^2 \mu_4 \\ -\sigma_{23}\sigma_{34} w_4^3 \\ +\sigma_{24}\sigma_{43} w_4^2 \mu_3 \\ +\sigma_{24}\sigma_{33} w_4^3 \\ -\sigma_{24}\sigma_{33} w_4^2 \mu_4 \end{array} \right) \right) \\
&+ \frac{1}{\sigma_{33}\sigma_{44} - \sigma_{34}\sigma_{43}} \left(\begin{array}{l} \sigma_{23}\sigma_{44} w_4 \\ \\ -\sigma_{24}\sigma_{43} w_4 \end{array} \right) \\
&\left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43} + \left(\mu_3 + \tfrac{\sigma_{34}}{\sigma_{44}}\left(w_4 - \mu_4\right)\right)^2\right)
\end{aligned}
\tag{A.37}
$$

We can invoke (A.11) directly to find the expectation of the first two terms in (A.37). We will now focus on the last additive term of (A.37). We simplify it as in

(A.30).

$$
\begin{array}{l}
\frac{1}{\sigma_{33}\sigma_{44}-\sigma_{34}\sigma_{43}}\begin{pmatrix}\sigma_{23}\sigma_{44}w_4\\ -\sigma_{24}\sigma_{43}w_4\end{pmatrix}\\
\left(\sigma_{33}-\sigma_{34}\sigma_{44}^{-1}\sigma_{43}+\left(\mu_3+\frac{\sigma_{34}}{\sigma_{44}}\left(w_4-\mu_4\right)\right)^2\right)=\\
\frac{(\sigma_{23}\sigma_{44}-\sigma_{24}\sigma_{43})}{\sigma_{33}\sigma_{44}-\sigma_{34}\sigma_{43}}w_4\\
\begin{pmatrix}\sigma_{33}-\sigma_{34}\sigma_{44}^{-1}\sigma_{43}+\frac{\sigma_{34}^2}{\sigma_{44}^2}w_4^2+\\ 2w_4\frac{\sigma_{34}}{\sigma_{44}}\left(\mu_3-\frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)+\left(\mu_3-\frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^2\end{pmatrix}=\\
\frac{\sigma_{23}\sigma_{44}-\sigma_{24}\sigma_{43}}{\sigma_{33}\sigma_{44}-\sigma_{34}\sigma_{43}}\\
\begin{pmatrix}\left(\sigma_{33}-\sigma_{34}\sigma_{44}^{-1}\sigma_{43}\right)w_4+\frac{\sigma_{34}^2}{\sigma_{44}^2}w_4^3+\\ 2w_4^2\frac{\sigma_{34}}{\sigma_{44}}\left(\mu_3-\frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)+\left(\mu_3-\frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^2w_4\end{pmatrix}
\end{array}
\tag{A.38}
$$

Now we plug back into (A.37).

$$
\begin{aligned}
&\mathrm{E}\left[\mathrm{E}\left[w_4 w_3 w_2 | w_4, w_3\right] | w_4\right] = \\
&\left(\begin{array}{l} w_4\mu_2 \\ \\ \\ \\ \\ \\ \end{array} \right. \\
&\quad \left. + \frac{1}{\sigma_{33}\sigma_{44} - \sigma_{34}\sigma_{43}} \left(\begin{array}{l} -\sigma_{23}\sigma_{44} w_4 \mu_3 \\ +\sigma_{23}\sigma_{34} w_4 \mu_4 \\ -\sigma_{23}\sigma_{34} w_4^2 \\ +\sigma_{24}\sigma_{43} w_4 \mu_3 \\ +\sigma_{24}\sigma_{33} w_4^2 \\ -\sigma_{24}\sigma_{33} w_4 \mu_4 \end{array} \right) \right) \left(\mu_3 - \tfrac{\sigma_{34}}{\sigma_{44}}\mu_4\right) \\
&+ \tfrac{\sigma_{34}}{\sigma_{44}} \left(\begin{array}{l} w_4^2\mu_2 \\ \\ \\ \\ \\ \\ \end{array} \right. \\
&\quad \left. + \frac{1}{\sigma_{33}\sigma_{44} - \sigma_{34}\sigma_{43}} \left(\begin{array}{l} -\sigma_{23}\sigma_{44} w_4^2 \mu_3 \\ +\sigma_{23}\sigma_{34} w_4^2 \mu_4 \\ -\sigma_{23}\sigma_{34} w_4^3 \\ +\sigma_{24}\sigma_{43} w_4^2 \mu_3 \\ +\sigma_{24}\sigma_{33} w_4^3 \\ -\sigma_{24}\sigma_{33} w_4^2 \mu_4 \end{array} \right) \right) \\
&+ \tfrac{\sigma_{23}\sigma_{44} - \sigma_{24}\sigma_{43}}{\sigma_{33}\sigma_{44} - \sigma_{34}\sigma_{43}} \\
&\left(\begin{array}{l} \left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43}\right) w_4 + \frac{\sigma_{34}^2}{\sigma_{44}^2} w_4^3 + \\ 2 w_4^2 \frac{\sigma_{34}}{\sigma_{44}} \left(\mu_3 - \frac{\sigma_{34}}{\sigma_{44}}\mu_4\right) + \left(\mu_3 - \frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^2 w_4 \end{array} \right)
\end{aligned}
\tag{A.39}
$$

We use (A.11) directly to find the expectation of (A.39), and thus $\mathrm{E}\left[w_4 w_3 w_2\right]$. The calculation of our last third order product moment identity, $\mathrm{E}\left[w_4^2 w_3\right]$ is

trivial. We will use (A.13) and (A.11).

$$\begin{aligned}
&\mathrm{E}\left[w_4^2 w_3\right] = \\
&\mathrm{E}\left[\mathrm{E}\left[w_4^2 w_3 | w_4\right]\right] = \\
&\mathrm{E}\left[w_4^2\left(\mu_3 + \tfrac{\sigma_{34}}{\sigma_{44}}\left(w_4 - \mu_4\right)\right)\right] = \\
&\mu_3 \mathrm{E}\left[w_4^2\right] + \tfrac{\sigma_{34}}{\sigma_{44}}\mathrm{E}\left[w_4^3\right] - \mu_4 \tfrac{\sigma_{34}}{\sigma_{44}}\mathrm{E}\left[w_4^2\right] = \\
&\mu_3\left(\sigma_{44} + \mu_4^2\right) + \tfrac{\sigma_{34}}{\sigma_{44}}\left(\mu_4\left(\mu_4^2 + 3\sigma_{44}\right)\right) \\
&-\mu_4 \tfrac{\sigma_{34}}{\sigma_{44}}\left(\sigma_{44} + \mu_4^2\right)
\end{aligned} \tag{A.40}$$

We have one more product moment identity to find, $\mathrm{E}\left[w_4 w_3 w_2 w_1\right]$. This is going to be the most complicated identity we derive in this section. As before, we use iterated conditional expectations.

$$\begin{aligned}
&\mathrm{E}\left[w_4 w_3 w_2 w_1\right] = \\
&\mathrm{E}\left[\mathrm{E}\left[w_4 w_3 w_2 w_1 | w_4, w_3, w_2\right]\right] = \\
&\mathrm{E}\left[\mathrm{E}\left[\mathrm{E}\left[w_4 w_3 w_2 w_1 | w_4, w_3, w_2\right] | w_4, w_3\right] | w_4\right]]
\end{aligned} \tag{A.41}$$

We begin with the innermost conditional expectation.

$$\begin{aligned}
&\mathrm{E}\left[w_4 w_3 w_2 w_1 | w_4, w_3, w_2\right] = \\
&w_4 w_3 w_2 \left(\begin{array}{l} \mu_1 + \\ \\ \left[\begin{array}{ccc} \sigma_{12} & \sigma_{13} & \sigma_{14} \end{array}\right] \left[\begin{array}{ccc} \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{42} & \sigma_{43} & \sigma_{44} \end{array}\right]^{-1} \left(\left[\begin{array}{c} w_2 \\ w_3 \\ w_4 \end{array}\right] - \left[\begin{array}{c} \mu_2 \\ \mu_3 \\ \mu_4 \end{array}\right] \right) \end{array} \right) = \\
&w_4 w_3 w_2 \mu_1 + \\
&\left[\begin{array}{ccc} \sigma_{12} & \sigma_{13} & \sigma_{14} \end{array}\right] \left[\begin{array}{ccc} \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{42} & \sigma_{43} & \sigma_{44} \end{array}\right]^{-1} \left[\begin{array}{c} w_4 w_3 w_2^2 - w_4 w_3 w_2 \mu_2 \\ w_4 w_3^2 w_2 - w_4 w_3 w_2 \mu_3 \\ w_4^2 w_3 w_2 - w_4 w_3 w_2 \mu_4 \end{array}\right]
\end{aligned} \tag{A.42}$$

We will next calculate the inverse of the covariance matrix of w_2, w_3, and w_4.

Call the covariance matrix $\mathbf{\Sigma}^{(\mathbf{2},\mathbf{3},\mathbf{4})}$. We will denote its inverse as $\mathbf{\Sigma}^{-(\mathbf{2},\mathbf{3},\mathbf{4})}$.

$$\mathbf{\Sigma}^{-(\mathbf{2},\mathbf{3},\mathbf{4})} = \begin{bmatrix} \sigma_{22}^{-(2,3,4)} & \sigma_{23}^{-(2,3,4)} & \sigma_{24}^{-(2,3,4)} \\ \sigma_{32}^{-(2,3,4)} & \sigma_{33}^{-(2,3,4)} & \sigma_{34}^{-(2,3,4)} \\ \sigma_{42}^{-(2,3,4)} & \sigma_{43}^{-(2,3,4)} & \sigma_{44}^{-(2,3,4)} \end{bmatrix} \tag{A.43}$$

$$\begin{aligned}
\sigma_{22}^{-(2,3,4)} &= \left|\mathbf{\Sigma}^{(\mathbf{2},\mathbf{3},\mathbf{4})}\right|^{-1} \left(\sigma_{33}\sigma_{44} - \sigma_{34}\sigma_{43}\right) \\
\sigma_{23}^{-(2,3,4)} &= \sigma_{32}^{-(2,3,4)} = -\left|\mathbf{\Sigma}^{(\mathbf{2},\mathbf{3},\mathbf{4})}\right|^{-1} \left(\sigma_{32}\sigma_{44} - \sigma_{42}\sigma_{34}\right) \\
\sigma_{24}^{-(2,3,4)} &= \sigma_{42}^{-(2,3,4)} = \left|\mathbf{\Sigma}^{(\mathbf{2},\mathbf{3},\mathbf{4})}\right|^{-1} \left(\sigma_{32}\sigma_{43} - \sigma_{33}\sigma_{42}\right) \\
\sigma_{33}^{-(2,3,4)} &= \left|\mathbf{\Sigma}^{(\mathbf{2},\mathbf{3},\mathbf{4})}\right|^{-1} \left(\sigma_{22}\sigma_{44} - \sigma_{24}\sigma_{42}\right) \\
\sigma_{34}^{-(2,3,4)} &= \sigma_{43}^{-(2,3,4)} = -\left|\mathbf{\Sigma}^{(\mathbf{2},\mathbf{3},\mathbf{4})}\right|^{-1} \left(\sigma_{22}\sigma_{43} - \sigma_{23}\sigma_{42}\right) \\
\sigma_{44}^{-(2,3,4)} &= \left|\mathbf{\Sigma}^{(\mathbf{2},\mathbf{3},\mathbf{4})}\right|^{-1} \left(\sigma_{22}\sigma_{33} - \sigma_{23}\sigma_{32}\right)
\end{aligned} \tag{A.44}$$

The determinant of the covariance matrix of $\mathbf{\Sigma}^{(\mathbf{2},\mathbf{3},\mathbf{4})}$ is the following.

$$\begin{aligned}
\left|\mathbf{\Sigma}^{(\mathbf{2},\mathbf{3},\mathbf{4})}\right| &= \begin{vmatrix} \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{42} & \sigma_{43} & \sigma_{44} \end{vmatrix} = \\
&\sigma_{22} \begin{vmatrix} \sigma_{33} & \sigma_{34} \\ \sigma_{43} & \sigma_{44} \end{vmatrix} - \sigma_{23} \begin{vmatrix} \sigma_{32} & \sigma_{34} \\ \sigma_{42} & \sigma_{44} \end{vmatrix} + \sigma_{24} \begin{vmatrix} \sigma_{32} & \sigma_{33} \\ \sigma_{42} & \sigma_{43} \end{vmatrix} = \\
&\sigma_{22}\left(\sigma_{33}\sigma_{44} - \sigma_{34}\sigma_{43}\right) \\
&-\sigma_{23}\left(\sigma_{32}\sigma_{44} - \sigma_{34}\sigma_{42}\right) \\
&+\sigma_{24}\left(\sigma_{32}\sigma_{43} - \sigma_{33}\sigma_{42}\right)
\end{aligned} \tag{A.45}$$

Now we return to (A.42).

$$
\begin{aligned}
&\mathrm{E}\left[w_4w_3w_2w_1|w_4,w_3,w_2\right] = \\
&w_4w_3w_2\mu_1+ \\
&\left[\begin{array}{ccc}\sigma_{12} & \sigma_{13} & \sigma_{14}\end{array}\right]\left[\begin{array}{ccc}\sigma_{22}^{-(2,3,4)} & \sigma_{23}^{-(2,3,4)} & \sigma_{24}^{-(2,3,4)}\\ \sigma_{32}^{-(2,3,4)} & \sigma_{33}^{-(2,3,4)} & \sigma_{34}^{-(2,3,4)}\\ \sigma_{42}^{-(2,3,4)} & \sigma_{43}^{-(2,3,4)} & \sigma_{44}^{-(2,3,4)}\end{array}\right]\left[\begin{array}{c} w_4w_3w_2^2 - w_4w_3w_2\mu_2\\ w_4w_3^2w_2 - w_4w_3w_2\mu_3\\ w_4^2w_3w_2 - w_4w_3w_2\mu_4\end{array}\right] = \\
&w_4w_3w_2\mu_1+ \\
&\left[\begin{array}{c}\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\\ \sigma_{12}\sigma_{23}^{-(2,3,4)} + \sigma_{13}\sigma_{33}^{-(2,3,4)} + \sigma_{14}\sigma_{43}^{-(2,3,4)}\\ \sigma_{12}\sigma_{24}^{-(2,3,4)} + \sigma_{13}\sigma_{34}^{-(2,3,4)} + \sigma_{14}\sigma_{44}^{-(2,3,4)}\end{array}\right]'\left[\begin{array}{c} w_4w_3w_2^2 - w_4w_3w_2\mu_2\\ w_4w_3^2w_2 - w_4w_3w_2\mu_3\\ w_4^2w_3w_2 - w_4w_3w_2\mu_4\end{array}\right] = \qquad \text{(A.46)} \\
&w_4w_3w_2\mu_1+ \\
&\left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right)\left(w_4w_3w_2^2 - w_4w_3w_2\mu_2\right) + \\
&\left(\sigma_{12}\sigma_{23}^{-(2,3,4)} + \sigma_{13}\sigma_{33}^{-(2,3,4)} + \sigma_{14}\sigma_{43}^{-(2,3,4)}\right)\left(w_4w_3^2w_2 - w_4w_3w_2\mu_3\right) + \\
&\left(\sigma_{12}\sigma_{24}^{-(2,3,4)} + \sigma_{13}\sigma_{34}^{-(2,3,4)} + \sigma_{14}\sigma_{44}^{-(2,3,4)}\right)\left(w_4^2w_3w_2 - w_4w_3w_2\mu_4\right)
\end{aligned}
$$

We will re-arrange (A.46) to clearly identify the powers of w_2 and their factors.

$$
\begin{aligned}
&\mathrm{E}\left[w_4w_3w_2w_1|w_4,w_3,w_2\right] = \\
&\left(\begin{array}{l} w_4w_3\mu_1\\ -\left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right)w_4w_3\mu_2\\ +\left(\sigma_{12}\sigma_{23}^{-(2,3,4)} + \sigma_{13}\sigma_{33}^{-(2,3,4)} + \sigma_{14}\sigma_{43}^{-(2,3,4)}\right)\left(w_4w_3^2 - w_4w_3\mu_3\right)\\ +\left(\sigma_{12}\sigma_{24}^{-(2,3,4)} + \sigma_{13}\sigma_{34}^{-(2,3,4)} + \sigma_{14}\sigma_{44}^{-(2,3,4)}\right)\left(w_4^2w_3 - w_4w_3\mu_4\right)\end{array}\right)w_2 \qquad \text{(A.47)} \\
&+\left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right)\left(w_4w_3\right)w_2^2
\end{aligned}
$$

Now we use this and (A.2) to obtain $\mathrm{E}\left[\mathrm{E}\left[w_4 w_3 w_2 w_1 | w_4, w_3, w_2\right] | w_4, w_3\right]$.

$$
\begin{aligned}
&\mathrm{E}\left[\mathrm{E}\left[w_4 w_3 w_2 w_1 | w_4, w_3, w_2\right] | w_4, w_3\right] = \\
&\left(\begin{array}{l}
w_4 w_3 \mu_1 \\
-\left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right) w_4 w_3 \mu_2 \\
+\left(\sigma_{12}\sigma_{23}^{-(2,3,4)} + \sigma_{13}\sigma_{33}^{-(2,3,4)} + \sigma_{14}\sigma_{43}^{-(2,3,4)}\right) \left(w_4 w_3^2 - w_4 w_3 \mu_3\right) \\
+\left(\sigma_{12}\sigma_{24}^{-(2,3,4)} + \sigma_{13}\sigma_{34}^{-(2,3,4)} + \sigma_{14}\sigma_{44}^{-(2,3,4)}\right) \left(w_4^2 w_3 - w_4 w_3 \mu_4\right)
\end{array} \right) \\
&\left(\mu_2 + \left[\begin{array}{cc} \sigma_{23} & \sigma_{24} \end{array} \right] \left[\begin{array}{cc} \sigma_{33} & \sigma_{34} \\ \sigma_{43} & \sigma_{44} \end{array} \right]^{-1} \left(\left[\begin{array}{c} w_3 \\ w_4 \end{array} \right] - \left[\begin{array}{c} \mu_3 \\ \mu_4 \end{array} \right] \right) \right) \\
&+\left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right) \left(w_4 w_3\right) \\
&\left(\begin{array}{l}
\sigma_{22} - \left[\begin{array}{cc} \sigma_{23} & \sigma_{24} \end{array} \right] \left[\begin{array}{cc} \sigma_{33} & \sigma_{34} \\ \sigma_{43} & \sigma_{44} \end{array} \right]^{-1} \left[\begin{array}{c} \sigma_{32} \\ \sigma_{42} \end{array} \right] \\
+\left(\mu_2 + \left[\begin{array}{cc} \sigma_{23} & \sigma_{24} \end{array} \right] \left[\begin{array}{cc} \sigma_{33} & \sigma_{34} \\ \sigma_{43} & \sigma_{44} \end{array} \right]^{-1} \left(\left[\begin{array}{c} w_3 \\ w_4 \end{array} \right] - \left[\begin{array}{c} \mu_3 \\ \mu_4 \end{array} \right] \right) \right)^2
\end{array} \right)
\end{aligned}
\tag{A.48}
$$

We need to re-arrange this expression so that the powers of w_3 and their factors are clearly identified. We focus on the first additive term, and utilize our matrix inverse notation from (A.44).

$$
\left[\begin{array}{cc} \sigma_{33} & \sigma_{34} \\ \sigma_{43} & \sigma_{44} \end{array} \right]^{-1} = \frac{1}{\sigma_{33}\sigma_{44} - \sigma_{34}\sigma_{43}} \left[\begin{array}{cc} \sigma_{44} & -\sigma_{34} \\ -\sigma_{43} & \sigma_{33} \end{array} \right] = \left[\begin{array}{cc} \sigma_{33}^{-(3,4)} & \sigma_{34}^{-(3,4)} \\ \sigma_{43}^{-(3,4)} & \sigma_{44}^{-(3,4)} \end{array} \right] \tag{A.49}
$$

$$
\begin{pmatrix}
\begin{pmatrix}
w_4 w_3 \mu_1 \\
-\left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right) w_4 w_3 \mu_2 \\
+\left(\sigma_{12}\sigma_{23}^{-(2,3,4)} + \sigma_{13}\sigma_{33}^{-(2,3,4)} + \sigma_{14}\sigma_{43}^{-(2,3,4)}\right) (w_4 w_3^2 - w_4 w_3 \mu_3) \\
+\left(\sigma_{12}\sigma_{24}^{-(2,3,4)} + \sigma_{13}\sigma_{34}^{-(2,3,4)} + \sigma_{14}\sigma_{44}^{-(2,3,4)}\right) (w_4^2 w_3 - w_4 w_3 \mu_4)
\end{pmatrix} \\
\left(\mu_2 + \begin{bmatrix} \sigma_{23} & \sigma_{24} \end{bmatrix} \begin{bmatrix} \sigma_{33} & \sigma_{34} \\ \sigma_{43} & \sigma_{44} \end{bmatrix}^{-1} \left(\begin{bmatrix} w_3 \\ w_4 \end{bmatrix} - \begin{bmatrix} \mu_3 \\ \mu_4 \end{bmatrix} \right) \right) =
\end{pmatrix}
$$

$$
\begin{pmatrix}
\begin{pmatrix}
w_4 \mu_1 \\
-\left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right) w_4 \mu_2 \\
-\left(\sigma_{12}\sigma_{23}^{-(2,3,4)} + \sigma_{13}\sigma_{33}^{-(2,3,4)} + \sigma_{14}\sigma_{43}^{-(2,3,4)}\right) w_4 \mu_3 \\
+\left(\sigma_{12}\sigma_{24}^{-(2,3,4)} + \sigma_{13}\sigma_{34}^{-(2,3,4)} + \sigma_{14}\sigma_{44}^{-(2,3,4)}\right) (w_4^2 - w_4 \mu_4)
\end{pmatrix} w_3 \\
+\left(\sigma_{12}\sigma_{23}^{-(2,3,4)} + \sigma_{13}\sigma_{33}^{-(2,3,4)} + \sigma_{14}\sigma_{43}^{-(2,3,4)}\right) w_4 w_3^2 \\
\left(\mu_2 + \begin{bmatrix} \sigma_{23} & \sigma_{24} \end{bmatrix} \begin{bmatrix} \sigma_{33}^{-(3,4)} & \sigma_{34}^{-(3,4)} \\ \sigma_{43}^{-(3,4)} & \sigma_{44}^{-(3,4)} \end{bmatrix} \begin{bmatrix} w_3 - \mu_3 \\ w_4 - \mu_4 \end{bmatrix}\right)
\end{pmatrix}
\tag{A.50}
$$

Now we will perform the matrix multiplications in and expand out the last multiplicative term.

$$
\begin{aligned}
&\mu_2 + \begin{bmatrix} \sigma_{23} & \sigma_{24} \end{bmatrix} \begin{bmatrix} \sigma_{33}^{-(3,4)} & \sigma_{34}^{-(3,4)} \\ \sigma_{43}^{-(3,4)} & \sigma_{44}^{-(3,4)} \end{bmatrix} \begin{bmatrix} w_3 - \mu_3 \\ w_4 - \mu_4 \end{bmatrix} = \\
&\mu_2 + \begin{bmatrix} \sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)} \\ \sigma_{23}\sigma_{34}^{-(3,4)} + \sigma_{24}\sigma_{44}^{-(3,4)} \end{bmatrix}' \begin{bmatrix} w_3 - \mu_3 \\ w_4 - \mu_4 \end{bmatrix} = \\
&\left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right) w_3 + \\
&\mu_2 - \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right) \mu_3 \\
&+\left(\sigma_{23}\sigma_{34}^{-(3,4)} + \sigma_{24}\sigma_{44}^{-(3,4)}\right) (w_4 - \mu_4)
\end{aligned}
\tag{A.51}
$$

We then return to (A.50) utilizing this work.

$$
\begin{aligned}
&\left(\begin{array}{l}
\left(\begin{array}{l}
w_4\mu_1 \\
-\left(\sigma_{12}\sigma_{22}^{-(2,3,4)}+\sigma_{13}\sigma_{32}^{-(2,3,4)}+\sigma_{14}\sigma_{42}^{-(2,3,4)}\right)w_4\mu_2 \\
-\left(\sigma_{12}\sigma_{23}^{-(2,3,4)}+\sigma_{13}\sigma_{33}^{-(2,3,4)}+\sigma_{14}\sigma_{43}^{-(2,3,4)}\right)w_4\mu_3 \\
+\left(\sigma_{12}\sigma_{24}^{-(2,3,4)}+\sigma_{13}\sigma_{34}^{-(2,3,4)}+\sigma_{14}\sigma_{44}^{-(2,3,4)}\right)(w_4^2-w_4\mu_4)
\end{array}\right)w_3 \\
+\left(\sigma_{12}\sigma_{23}^{-(2,3,4)}+\sigma_{13}\sigma_{33}^{-(2,3,4)}+\sigma_{14}\sigma_{43}^{-(2,3,4)}\right)w_4w_3^2
\end{array}\right) \\
&\left(\begin{array}{l}
\left(\sigma_{23}\sigma_{33}^{-(3,4)}+\sigma_{24}\sigma_{43}^{-(3,4)}\right)w_3+ \\
\mu_2-\left(\sigma_{23}\sigma_{33}^{-(3,4)}+\sigma_{24}\sigma_{43}^{-(3,4)}\right)\mu_3 \\
+\left(\sigma_{23}\sigma_{34}^{-(3,4)}+\sigma_{24}\sigma_{44}^{-(3,4)}\right)(w_4-\mu_4)
\end{array}\right)= \\
&\left(\begin{array}{l}
\left(\sigma_{23}\sigma_{33}^{-(3,4)}+\sigma_{24}\sigma_{43}^{-(3,4)}\right) \\
\left(\begin{array}{l}
w_4\mu_1 \\
-\left(\sigma_{12}\sigma_{22}^{-(2,3,4)}+\sigma_{13}\sigma_{32}^{-(2,3,4)}+\sigma_{14}\sigma_{42}^{-(2,3,4)}\right)w_4\mu_2 \\
-\left(\sigma_{12}\sigma_{23}^{-(2,3,4)}+\sigma_{13}\sigma_{33}^{-(2,3,4)}+\sigma_{14}\sigma_{43}^{-(2,3,4)}\right)w_4\mu_3 \\
+\left(\sigma_{12}\sigma_{24}^{-(2,3,4)}+\sigma_{13}\sigma_{34}^{-(2,3,4)}+\sigma_{14}\sigma_{44}^{-(2,3,4)}\right)(w_4^2-w_4\mu_4)
\end{array}\right)
\end{array}\right)w_3^2+ \\
&\left(\begin{array}{l}
\left(\sigma_{23}\sigma_{33}^{-(3,4)}+\sigma_{24}\sigma_{43}^{-(3,4)}\right) \\
\left(\sigma_{12}\sigma_{23}^{-(2,3,4)}+\sigma_{13}\sigma_{33}^{-(2,3,4)}+\sigma_{14}\sigma_{43}^{-(2,3,4)}\right)w_4
\end{array}\right)w_3^3+ \\
&\left(\begin{array}{l}
\mu_2-\left(\sigma_{23}\sigma_{33}^{-(3,4)}+\sigma_{24}\sigma_{43}^{-(3,4)}\right)\mu_3 \\
+\left(\sigma_{23}\sigma_{34}^{-(3,4)}+\sigma_{24}\sigma_{44}^{-(3,4)}\right)(w_4-\mu_4)
\end{array}\right) \\
&\left(\begin{array}{l}
\left(\begin{array}{l}
w_4\mu_1 \\
-\left(\sigma_{12}\sigma_{22}^{-(2,3,4)}+\sigma_{13}\sigma_{32}^{-(2,3,4)}+\sigma_{14}\sigma_{42}^{-(2,3,4)}\right)w_4\mu_2 \\
-\left(\sigma_{12}\sigma_{23}^{-(2,3,4)}+\sigma_{13}\sigma_{33}^{-(2,3,4)}+\sigma_{14}\sigma_{43}^{-(2,3,4)}\right)w_4\mu_3 \\
+\left(\sigma_{12}\sigma_{24}^{-(2,3,4)}+\sigma_{13}\sigma_{34}^{-(2,3,4)}+\sigma_{14}\sigma_{44}^{-(2,3,4)}\right)(w_4^2-w_4\mu_4)
\end{array}\right)w_3 \\
+\left(\sigma_{12}\sigma_{23}^{-(2,3,4)}+\sigma_{13}\sigma_{33}^{-(2,3,4)}+\sigma_{14}\sigma_{43}^{-(2,3,4)}\right)w_4w_3^2
\end{array}\right)
\end{aligned}
\tag{A.52}
$$

So (A.52) gives us the first additive term of (A.48). Now we will focus on the

second. We immediately take advantage of (A.51).

$$
\begin{array}{l}
\left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right)(w_4 w_3) \\
\left(\begin{array}{l}
\sigma_{22} - \begin{bmatrix} \sigma_{23} & \sigma_{24} \end{bmatrix} \begin{bmatrix} \sigma_{33} & \sigma_{34} \\ \sigma_{43} & \sigma_{44} \end{bmatrix}^{-1} \begin{bmatrix} \sigma_{32} \\ \sigma_{42} \end{bmatrix} \\
+ \left(\mu_2 + \begin{bmatrix} \sigma_{23} & \sigma_{24} \end{bmatrix} \begin{bmatrix} \sigma_{33} & \sigma_{34} \\ \sigma_{43} & \sigma_{44} \end{bmatrix}^{-1} \left(\begin{bmatrix} w_3 \\ w_4 \end{bmatrix} - \begin{bmatrix} \mu_3 \\ \mu_4 \end{bmatrix}\right)\right)^2
\end{array}\right) = \\
\left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right)(w_4 w_3) \\
\left(\begin{array}{l}
\sigma_{22} - \begin{bmatrix} \sigma_{23} & \sigma_{24} \end{bmatrix} \begin{bmatrix} \sigma_{33} & \sigma_{34} \\ \sigma_{43} & \sigma_{44} \end{bmatrix}^{-1} \begin{bmatrix} \sigma_{32} \\ \sigma_{42} \end{bmatrix} \\
+ \left(\begin{array}{l}
\left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right) w_3 + \\
\left(\begin{array}{l}
\mu_2 - \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right)\mu_3 \\
+ \left(\sigma_{23}\sigma_{34}^{-(3,4)} + \sigma_{24}\sigma_{44}^{-(3,4)}\right)(w_4 - \mu_4)
\end{array}\right)
\end{array}\right)^2
\end{array}\right) = \\
\left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right)(w_4 w_3) \\
\left(\begin{array}{l}
\sigma_{22} - \begin{bmatrix} \sigma_{23} & \sigma_{24} \end{bmatrix} \begin{bmatrix} \sigma_{33} & \sigma_{34} \\ \sigma_{43} & \sigma_{44} \end{bmatrix}^{-1} \begin{bmatrix} \sigma_{32} \\ \sigma_{42} \end{bmatrix} \\
+ \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right)^2 w_3^2 + \\
2\left(\begin{array}{l}
\mu_2 - \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right)\mu_3 \\
+ \left(\sigma_{23}\sigma_{34}^{-(3,4)} + \sigma_{24}\sigma_{44}^{-(3,4)}\right)(w_4 - \mu_4)
\end{array}\right) \\
\left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right) w_3 + \\
\left(\begin{array}{l}
\mu_2 - \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right)\mu_3 \\
+ \left(\sigma_{23}\sigma_{34}^{-(3,4)} + \sigma_{24}\sigma_{44}^{-(3,4)}\right)(w_4 - \mu_4)
\end{array}\right)^2
\end{array}\right)
\end{array}
\tag{A.53}
$$

Now we will multiply through the outermost w_3 term.

$$
\begin{array}{l}
\left(\sigma_{12}\sigma_{22}^{-(2,3,4)}+\sigma_{13}\sigma_{32}^{-(2,3,4)}+\sigma_{14}\sigma_{42}^{-(2,3,4)}\right)w_4 \\
\left(\begin{array}{l}
\left(\sigma_{22}-\begin{bmatrix}\sigma_{23} & \sigma_{24}\end{bmatrix}\begin{bmatrix}\sigma_{33} & \sigma_{34}\\ \sigma_{43} & \sigma_{44}\end{bmatrix}^{-1}\begin{bmatrix}\sigma_{32}\\ \sigma_{42}\end{bmatrix}\right)w_3 \\
+\left(\sigma_{23}\sigma_{33}^{-(3,4)}+\sigma_{24}\sigma_{43}^{-(3,4)}\right)^2 w_3^3+ \\
2\left(\begin{array}{l}\mu_2-\left(\sigma_{23}\sigma_{33}^{-(3,4)}+\sigma_{24}\sigma_{43}^{-(3,4)}\right)\mu_3 \\ +\left(\sigma_{23}\sigma_{34}^{-(3,4)}+\sigma_{24}\sigma_{44}^{-(3,4)}\right)(w_4-\mu_4)\end{array}\right) \\
\left(\sigma_{23}\sigma_{33}^{-(3,4)}+\sigma_{24}\sigma_{43}^{-(3,4)}\right)w_3^2+ \\
\left(\begin{array}{l}\mu_2-\left(\sigma_{23}\sigma_{33}^{-(3,4)}+\sigma_{24}\sigma_{43}^{-(3,4)}\right)\mu_3 \\ +\left(\sigma_{23}\sigma_{34}^{-(3,4)}+\sigma_{24}\sigma_{44}^{-(3,4)}\right)(w_4-\mu_4)\end{array}\right)^2 w_3
\end{array}\right)= \\
\left(\begin{array}{l}\left(\sigma_{12}\sigma_{22}^{-(2,3,4)}+\sigma_{13}\sigma_{32}^{-(2,3,4)}+\sigma_{14}\sigma_{42}^{-(2,3,4)}\right)w_4 \\ \left(\sigma_{23}\sigma_{33}^{-(3,4)}+\sigma_{24}\sigma_{43}^{-(3,4)}\right)^2\end{array}\right)w_3^3+ \\
\left(\begin{array}{l}\left(\sigma_{12}\sigma_{22}^{-(2,3,4)}+\sigma_{13}\sigma_{32}^{-(2,3,4)}+\sigma_{14}\sigma_{42}^{-(2,3,4)}\right)w_4 \\ 2\left(\begin{array}{l}\mu_2-\left(\sigma_{23}\sigma_{33}^{-(3,4)}+\sigma_{24}\sigma_{43}^{-(3,4)}\right)\mu_3 \\ +\left(\sigma_{23}\sigma_{34}^{-(3,4)}+\sigma_{24}\sigma_{44}^{-(3,4)}\right)(w_4-\mu_4)\end{array}\right) \\ \left(\sigma_{23}\sigma_{33}^{-(3,4)}+\sigma_{24}\sigma_{43}^{-(3,4)}\right)\end{array}\right)w_3^2+ \\
\left(\begin{array}{l}\left(\sigma_{12}\sigma_{22}^{-(2,3,4)}+\sigma_{13}\sigma_{32}^{-(2,3,4)}+\sigma_{14}\sigma_{42}^{-(2,3,4)}\right)w_4 \\ \left(\begin{array}{l}\left(\sigma_{22}-\begin{bmatrix}\sigma_{23} & \sigma_{24}\end{bmatrix}\begin{bmatrix}\sigma_{33} & \sigma_{34}\\ \sigma_{43} & \sigma_{44}\end{bmatrix}^{-1}\begin{bmatrix}\sigma_{32}\\ \sigma_{42}\end{bmatrix}\right) \\ +\left(\begin{array}{l}\mu_2-\left(\sigma_{23}\sigma_{33}^{-(3,4)}+\sigma_{24}\sigma_{43}^{-(3,4)}\right)\mu_3 \\ +\left(\sigma_{23}\sigma_{34}^{-(3,4)}+\sigma_{24}\sigma_{44}^{-(3,4)}\right)(w_4-\mu_4)\end{array}\right)^2\end{array}\right)w_3\end{array}\right)
\end{array}
\tag{A.54}
$$

There is an incomplete matrix inversion and matrix multiplication in (A.54). We

will calculate the involved term and then plug it back into (A.54).

$$\begin{aligned}
&\begin{bmatrix} \sigma_{23} & \sigma_{24} \end{bmatrix} \begin{bmatrix} \sigma_{33} & \sigma_{34} \\ \sigma_{43} & \sigma_{44} \end{bmatrix}^{-1} \begin{bmatrix} \sigma_{32} \\ \sigma_{42} \end{bmatrix} = \\
&\begin{bmatrix} \sigma_{23} & \sigma_{24} \end{bmatrix} \begin{bmatrix} \sigma_{33}^{-(3,4)} & \sigma_{34}^{-(3,4)} \\ \sigma_{43}^{-(3,4)} & \sigma_{44}^{-(3,4)} \end{bmatrix} \begin{bmatrix} \sigma_{32} \\ \sigma_{42} \end{bmatrix} = \\
&\begin{bmatrix} \sigma_{23} & \sigma_{24} \end{bmatrix} \begin{bmatrix} \sigma_{33}^{-(3,4)}\sigma_{32} + \sigma_{34}^{-(3,4)}\sigma_{42} \\ \sigma_{43}^{-(3,4)}\sigma_{32} + \sigma_{44}^{-(3,4)}\sigma_{42} \end{bmatrix} = \\
&\sigma_{23}\left(\sigma_{33}^{-(3,4)}\sigma_{32} + \sigma_{34}^{-(3,4)}\sigma_{42}\right) + \sigma_{24}\left(\sigma_{43}^{-(3,4)}\sigma_{32} + \sigma_{44}^{-(3,4)}\sigma_{42}\right)
\end{aligned} \tag{A.55}$$

Now we plug this into (A.54).

$$\begin{aligned}
&\left(\begin{array}{l} \left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right) w_4 \\ \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right)^2 \end{array} \right) w_3^3 + \\
&\left(\begin{array}{l} \left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right) w_4 \\ 2\left(\begin{array}{l} \mu_2 - \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right)\mu_3 \\ + \left(\sigma_{23}\sigma_{34}^{-(3,4)} + \sigma_{24}\sigma_{44}^{-(3,4)}\right)(w_4 - \mu_4) \end{array} \right) \\ \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right) \end{array} \right) w_3^2 + \\
&\left(\begin{array}{l} \left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right) w_4 \\ \left(\begin{array}{l} \left(\sigma_{22} - \left(\begin{array}{l} \sigma_{23}\left(\sigma_{33}^{-(3,4)}\sigma_{32} + \sigma_{34}^{-(3,4)}\sigma_{42}\right) + \\ \sigma_{24}\left(\sigma_{43}^{-(3,4)}\sigma_{32} + \sigma_{44}^{-(3,4)}\sigma_{42}\right) \end{array} \right) \right)^2 \\ + \left(\begin{array}{l} \mu_2 - \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right)\mu_3 \\ + \left(\sigma_{23}\sigma_{34}^{-(3,4)} + \sigma_{24}\sigma_{44}^{-(3,4)}\right)(w_4 - \mu_4) \end{array} \right)^2 \end{array} \right) \end{array} \right) w_3
\end{aligned} \tag{A.56}$$

Now we use (A.56) and and (A.52) to write, with multiples of w_3 powers clearly identified,$\mathrm{E}\left[\mathrm{E}\left[w_4w_3w_2w_1|w_4,w_3,w_2\right]|w_4,w_3\right]$.

$$\mathrm{E}\left[\mathrm{E}\left[w_4w_3w_2w_1|w_4,w_3,w_2\right]|w_4,w_3\right] = aw_3^3 + bw_3^2 + cw_3 \tag{A.57}$$

The conditional expectation has the following cubic coefficient.

$$
\begin{aligned}
a = & \\
& \begin{pmatrix} \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right) \\ \left(\sigma_{12}\sigma_{23}^{-(2,3,4)} + \sigma_{13}\sigma_{33}^{-(2,3,4)} + \sigma_{14}\sigma_{43}^{-(2,3,4)}\right) w_4 \end{pmatrix} \\
+ & \begin{pmatrix} \left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right) w_4 \\ \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right)^2 \end{pmatrix} = \\
& \left(\begin{Bmatrix} \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right) \\ \left(\sigma_{12}\sigma_{23}^{-(2,3,4)} + \sigma_{13}\sigma_{33}^{-(2,3,4)} + \sigma_{14}\sigma_{43}^{-(2,3,4)}\right) \end{Bmatrix} + \begin{Bmatrix} \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right)^2 \\ \left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right) \end{Bmatrix} \right) w_4
\end{aligned}
\tag{A.58}
$$

The quadratic coefficient is given below.

$$
\begin{aligned}
b = & \left(\begin{array}{l}
\left(\begin{array}{l} \left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right) w_4 \\ 2\left(\begin{array}{l} \mu_2 - \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right)\mu_3 \\ + \left(\sigma_{23}\sigma_{34}^{-(3,4)} + \sigma_{24}\sigma_{44}^{-(3,4)}\right)(w_4 - \mu_4) \end{array} \right) \\ \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right) \end{array} \right) + \\
\left(\begin{array}{l} \left(\begin{array}{l} \mu_2 - \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right)\mu_3 \\ + \left(\sigma_{23}\sigma_{34}^{-(3,4)} + \sigma_{24}\sigma_{44}^{-(3,4)}\right)(w_4 - \mu_4) \end{array} \right) \\ \left(\sigma_{12}\sigma_{23}^{-(2,3,4)} + \sigma_{13}\sigma_{33}^{-(2,3,4)} + \sigma_{14}\sigma_{43}^{-(2,3,4)}\right) w_4 \end{array} \right) + \\
\left(\begin{array}{l} \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right) \\ \left(\begin{array}{l} w_4\mu_1 \\ - \left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right) w_4\mu_2 \\ - \left(\sigma_{12}\sigma_{23}^{-(2,3,4)} + \sigma_{13}\sigma_{33}^{-(2,3,4)} + \sigma_{14}\sigma_{43}^{-(2,3,4)}\right) w_4\mu_3 \\ + \left(\sigma_{12}\sigma_{24}^{-(2,3,4)} + \sigma_{13}\sigma_{34}^{-(2,3,4)} + \sigma_{14}\sigma_{44}^{-(2,3,4)}\right)(w_4^2 - w_4\mu_4) \end{array} \right) \end{array} \right)
\end{array} \right)
\end{aligned}
\tag{A.59}
$$

Finally, the linear coefficient is the following.

$$c = \left(\begin{array}{l} \left(\begin{array}{l} \left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)} \right) w_4 \\ \left(\begin{array}{l} \left(\sigma_{22} - \left(\begin{array}{l} \sigma_{23}\left(\sigma_{33}^{-(3,4)}\sigma_{32} + \sigma_{34}^{-(3,4)}\sigma_{42} \right) + \\ \sigma_{24}\left(\sigma_{43}^{-(3,4)}\sigma_{32} + \sigma_{44}^{-(3,4)}\sigma_{42} \right) \end{array} \right) \right)^2 \\ + \left(\begin{array}{l} \mu_2 - \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)} \right) \mu_3 \\ + \left(\sigma_{23}\sigma_{34}^{-(3,4)} + \sigma_{24}\sigma_{44}^{-(3,4)} \right) (w_4 - \mu_4) \end{array} \right)^2 \end{array} \right) \end{array} \right) + \\ \left(\begin{array}{l} \left(\begin{array}{l} \mu_2 - \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)} \right) \mu_3 \\ + \left(\sigma_{23}\sigma_{34}^{-(3,4)} + \sigma_{24}\sigma_{44}^{-(3,4)} \right) (w_4 - \mu_4) \end{array} \right) \\ \left(\begin{array}{l} w_4\mu_1 \\ - \left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)} \right) w_4\mu_2 \\ - \left(\sigma_{12}\sigma_{23}^{-(2,3,4)} + \sigma_{13}\sigma_{33}^{-(2,3,4)} + \sigma_{14}\sigma_{43}^{-(2,3,4)} \right) w_4\mu_3 \\ + \left(\sigma_{12}\sigma_{24}^{-(2,3,4)} + \sigma_{13}\sigma_{34}^{-(2,3,4)} + \sigma_{14}\sigma_{44}^{-(2,3,4)} \right) (w_4^2 - w_4\mu_4) \end{array} \right) \end{array} \right) \end{array} \right) \tag{A.60}$$

It is time to take our last conditional expectation. We will condition on w_4 and take the expectation of (A.57).

$$\begin{array}{l} \mathrm{E}\left[\mathrm{E}\left[\mathrm{E}\left[w_4w_3w_2w_1|w_4,w_3,w_2\right]|w_4,w_3\right]|w_4\right] = \\ a\mathrm{E}\left[w_3^3|w_4\right] + b\mathrm{E}\left[w_3^2|w_4\right] + c\mathrm{E}\left[w_3|w_4\right] \end{array} \tag{A.61}$$

To calculate the first additive term, we use (A.11) and (A.13).

$$\begin{array}{l} a\mathrm{E}\left[w_3^3|w_4\right] = \\ a\left(\mathrm{E}\left[w_3|w_4\right]\left(\mathrm{E}\left[w_3|w_4\right]^2 + 3\mathrm{Var}\left[w_3|w_4\right]\right)\right) = \\ a\left(\begin{array}{l} \left(\mu_3 + \frac{\sigma_{34}}{\sigma_{44}}\left(w_4 - \mu_4\right)\right) \\ \left(\left(\mu_3 + \frac{\sigma_{34}}{\sigma_{44}}\left(w_4 - \mu_4\right)\right)^2 + 3\left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43}\right)\right) \end{array} \right) \end{array} \tag{A.62}$$

The next term is more simply computed.

$$\begin{aligned} &b\mathrm{E}\left[w_3^2|w_4\right] = \\ &b\left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43} + \left(\mu_3 + \tfrac{\sigma_{34}}{\sigma_{44}}\left(w_4 - \mu_4\right)\right)^2\right) \end{aligned} \tag{A.63}$$

The last additive term is trivial to compute.

$$c\mathrm{E}\left[w_3|w_4\right] = c\left(\mu_3 + \frac{\sigma_{34}}{\sigma_{44}}\left(w_4 - \mu_4\right)\right) \tag{A.64}$$

To finally calculate $\mathrm{E}\left[w_4 w_3 w_2 w_1\right]$ we must take the expectation of (A.61). To do this we will take the expectations of (5.70) - (5.72) and add them together. We begin with the expectation of (A.62).

$$\begin{aligned} &\mathrm{E}\left[a\mathrm{E}\left[w_3^3|w_4\right]\right] = \\ &\mathrm{E}\left[\begin{array}{c} \left(\begin{array}{c} \left(\begin{array}{c} \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right) \\ \left(\sigma_{12}\sigma_{23}^{-(2,3,4)} + \sigma_{13}\sigma_{33}^{-(2,3,4)} + \sigma_{14}\sigma_{43}^{-(2,3,4)}\right) \end{array}\right) + \\ \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right)^2 \\ \left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right) \end{array}\right) w_4 \\ \left(\begin{array}{c} \left(\mu_3 + \frac{\sigma_{34}}{\sigma_{44}}\left(w_4 - \mu_4\right)\right) \\ \left(\left(\mu_3 + \frac{\sigma_{34}}{\sigma_{44}}\left(w_4 - \mu_4\right)\right)^2 + 3\left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43}\right)\right) \end{array}\right) \end{array}\right] \end{aligned} \tag{A.65}$$

Now we expand and group powers of w_4 in the second multiplicative term.

$$\mathrm{E}\left[\begin{array}{l}\left(\begin{array}{l}\left(\begin{array}{l}\left(\sigma_{23}\sigma_{33}^{-(3,4)}+\sigma_{24}\sigma_{43}^{-(3,4)}\right)\\ \left(\sigma_{12}\sigma_{23}^{-(2,3,4)}+\sigma_{13}\sigma_{33}^{-(2,3,4)}+\sigma_{14}\sigma_{43}^{-(2,3,4)}\right)\end{array}\right)+\\ \left(\begin{array}{l}\left(\sigma_{23}\sigma_{33}^{-(3,4)}+\sigma_{24}\sigma_{43}^{-(3,4)}\right)^2\\ \left(\sigma_{12}\sigma_{22}^{-(2,3,4)}+\sigma_{13}\sigma_{32}^{-(2,3,4)}+\sigma_{14}\sigma_{42}^{-(2,3,4)}\right)\end{array}\right)\end{array}\right)w_4\\ \left(\begin{array}{l}\left(\left(\mu_3-\frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)+\frac{\sigma_{34}}{\sigma_{44}}w_4\right)\\ \left(\begin{array}{l}\frac{\sigma_{34}^2}{\sigma_{44}^2}w_4^2+2\frac{\sigma_{34}}{\sigma_{44}}w_4\left(\mu_3-\frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)\\ +\left(\mu_3-\frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^2+3\left(\sigma_{33}-\sigma_{34}\sigma_{44}^{-1}\sigma_{43}\right)\end{array}\right)\end{array}\right)\end{array}\right] \tag{A.66}$$

Next we multiply through the outer w_4.

$$\mathrm{E}\left[\begin{array}{l}\left(\begin{array}{l}\left(\begin{array}{l}\left(\begin{array}{l}\left(\sigma_{23}\sigma_{33}^{-(3,4)}+\sigma_{24}\sigma_{43}^{-(3,4)}\right)\\ \left(\sigma_{12}\sigma_{23}^{-(2,3,4)}+\sigma_{13}\sigma_{33}^{-(2,3,4)}+\sigma_{14}\sigma_{43}^{-(2,3,4)}\right)\end{array}\right)+\\ \left(\begin{array}{l}\left(\sigma_{23}\sigma_{33}^{-(3,4)}+\sigma_{24}\sigma_{43}^{-(3,4)}\right)^2\\ \left(\sigma_{12}\sigma_{22}^{-(2,3,4)}+\sigma_{13}\sigma_{32}^{-(2,3,4)}+\sigma_{14}\sigma_{42}^{-(2,3,4)}\right)\end{array}\right)\end{array}\right)\\ \left(\mu_3-\frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)w_4\left(\begin{array}{l}\frac{\sigma_{34}^2}{\sigma_{44}^2}w_4^2+2\frac{\sigma_{34}}{\sigma_{44}}w_4\left(\mu_3-\frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)\\ +\left(\mu_3-\frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^2+3\left(\sigma_{33}-\sigma_{34}\sigma_{44}^{-1}\sigma_{43}\right)\end{array}\right)\end{array}\right)\\ +\left(\frac{\sigma_{34}}{\sigma_{44}}w_4^2\left(\begin{array}{l}\frac{\sigma_{34}^2}{\sigma_{44}^2}w_4^2+2\frac{\sigma_{34}}{\sigma_{44}}w_4\left(\mu_3-\frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)\\ +\left(\mu_3-\frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^2+3\left(\sigma_{33}-\sigma_{34}\sigma_{44}^{-1}\sigma_{43}\right)\end{array}\right)\right)\end{array}\right] \tag{A.67}$$

Now we multiply through the outer w_4 and w_4^2 terms.

$$\mathrm{E}\left[\begin{array}{l}\left(\begin{array}{l}\left(\begin{array}{l}\left(\sigma_{23}\sigma_{33}^{-(3,4)}+\sigma_{24}\sigma_{43}^{-(3,4)}\right)\\ \left(\sigma_{12}\sigma_{23}^{-(2,3,4)}+\sigma_{13}\sigma_{33}^{-(2,3,4)}+\sigma_{14}\sigma_{43}^{-(2,3,4)}\right)\end{array}\right)+\\ \left(\begin{array}{l}\left(\sigma_{23}\sigma_{33}^{-(3,4)}+\sigma_{24}\sigma_{43}^{-(3,4)}\right)^2\\ \left(\sigma_{12}\sigma_{22}^{-(2,3,4)}+\sigma_{13}\sigma_{32}^{-(2,3,4)}+\sigma_{14}\sigma_{42}^{-(2,3,4)}\right)\end{array}\right)\end{array}\right)\\ \left(\begin{array}{l}\left(\begin{array}{l}\frac{\sigma_{34}^2}{\sigma_{44}^2}w_4^3+2\frac{\sigma_{34}}{\sigma_{44}}w_4^2\left(\mu_3-\frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)\\ +\left(\mu_3-\frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^2w_4+\\ 3\left(\sigma_{33}-\sigma_{34}\sigma_{44}^{-1}\sigma_{43}\right)w_4\end{array}\right)\\ \left(\mu_3-\frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)\end{array}\right)\\ +\left(\frac{\sigma_{34}}{\sigma_{44}}\left(\begin{array}{l}\frac{\sigma_{34}^2}{\sigma_{44}^2}w_4^4+2\frac{\sigma_{34}}{\sigma_{44}}w_4^3\left(\mu_3-\frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)\\ +\left(\mu_3-\frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^2w_4^2+3\left(\sigma_{33}-\sigma_{34}\sigma_{44}^{-1}\sigma_{43}\right)w_4^2\end{array}\right)\right)\end{array}\right] \tag{A.68}$$

This expectation can be easily found by using (A.11) and (A.10). Now we will find the expectation of (A.63). First we will simplify b.

$$b=\left(\begin{array}{l}\left(\begin{array}{l}\left(\sigma_{12}\sigma_{22}^{-(2,3,4)}+\sigma_{13}\sigma_{32}^{-(2,3,4)}+\sigma_{14}\sigma_{42}^{-(2,3,4)}\right)\\ 2\left(\begin{array}{l}w_4\mu_2-\left(\sigma_{23}\sigma_{33}^{-(3,4)}+\sigma_{24}\sigma_{43}^{-(3,4)}\right)w_4\mu_3\\ +\left(\sigma_{23}\sigma_{34}^{-(3,4)}+\sigma_{24}\sigma_{44}^{-(3,4)}\right)(w_4^2-w_4\mu_4)\end{array}\right)+\\ \left(\sigma_{23}\sigma_{33}^{-(3,4)}+\sigma_{24}\sigma_{43}^{-(3,4)}\right)\end{array}\right)+\\ \left(\begin{array}{l}\left(\begin{array}{l}w_4\mu_2-\left(\sigma_{23}\sigma_{33}^{-(3,4)}+\sigma_{24}\sigma_{43}^{-(3,4)}\right)w_4\mu_3\\ +\left(\sigma_{23}\sigma_{34}^{-(3,4)}+\sigma_{24}\sigma_{44}^{-(3,4)}\right)(w_4^2-w_4\mu_4)\end{array}\right)+\\ \left(\sigma_{12}\sigma_{23}^{-(2,3,4)}+\sigma_{13}\sigma_{33}^{-(2,3,4)}+\sigma_{14}\sigma_{43}^{-(2,3,4)}\right)\end{array}\right)\\ \left(\begin{array}{l}\left(\sigma_{23}\sigma_{33}^{-(3,4)}+\sigma_{24}\sigma_{43}^{-(3,4)}\right)\\ \left(\begin{array}{l}w_4\mu_1\\ -\left(\sigma_{12}\sigma_{22}^{-(2,3,4)}+\sigma_{13}\sigma_{32}^{-(2,3,4)}+\sigma_{14}\sigma_{42}^{-(2,3,4)}\right)w_4\mu_2\\ -\left(\sigma_{12}\sigma_{23}^{-(2,3,4)}+\sigma_{13}\sigma_{33}^{-(2,3,4)}+\sigma_{14}\sigma_{43}^{-(2,3,4)}\right)w_4\mu_3\\ +\left(\sigma_{12}\sigma_{24}^{-(2,3,4)}+\sigma_{13}\sigma_{34}^{-(2,3,4)}+\sigma_{14}\sigma_{44}^{-(2,3,4)}\right)(w_4^2-w_4\mu_4)\end{array}\right)\end{array}\right)\end{array}\right) \tag{A.69}$$

Next we expand $\mathrm{E}\left[w_3^2|w_4\right]$.

$$\begin{array}{l}
\mathrm{E}\left[w_3^2|w_4\right] = \\
\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43} + \left(\frac{\sigma_{34}}{\sigma_{44}}w_4 + \left(\mu_3 - \frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)\right)^2 = \\
\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43} + \frac{\sigma_{34}^2}{\sigma_{44}^2}w_4^2 + \\
2\frac{\sigma_{34}}{\sigma_{44}}w_4\left(\mu_3 - \frac{\sigma_{34}}{\sigma_{44}}\mu_4\right) + \left(\mu_3 - \frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^2 = \\
\left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43} + \left(\mu_3 - \frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^2\right) \\
+\frac{\sigma_{34}^2}{\sigma_{44}^2}w_4^2 + 2\frac{\sigma_{34}}{\sigma_{44}}w_4\left(\mu_3 - \frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)
\end{array} \tag{A.70}$$

Now we will compute $\mathrm{E}\left[b\mathrm{E}\left[w_3^2|w_4\right]\right]$.

$$\begin{array}{l}
\mathrm{E}\left[b\mathrm{E}\left[w_3^2|w_4\right]\right] = \\
\mathrm{E}\left[b\left(\begin{array}{l}\left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43} + \left(\mu_3 - \frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^2\right) \\ +\frac{\sigma_{34}^2}{\sigma_{44}^2}w_4^2 + 2\frac{\sigma_{34}}{\sigma_{44}}w_4\left(\mu_3 - \frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)\end{array}\right)\right] = \\
\left(\sigma_{33} - \sigma_{34}\sigma_{44}^{-1}\sigma_{43} + \left(\mu_3 - \frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)^2\right)\mathrm{E}\left[b\right] + \\
\frac{\sigma_{34}^2}{\sigma_{44}^2}\mathrm{E}\left[w_4^2 b\right] + 2\frac{\sigma_{34}}{\sigma_{44}}\left(\mu_3 - \frac{\sigma_{34}}{\sigma_{44}}\mu_4\right)\mathrm{E}\left[w_4 b\right]
\end{array} \tag{A.71}$$

The expectation $\mathrm{E}\left[b\right]$ is easily calculated by elementary principles. The expectations $\mathrm{E}\left[w_4^2 b\right]$ and $\mathrm{E}\left[w_4 b\right]$ are easily calculated by using (5.19) and (5.18). We will multiply through the powers of w_4 in both the $w_4^2 b$ and $w_4 b$ expressions to demonstrate

this.

$$
\begin{aligned}
w_4^2 b = \\
&\left(\begin{array}{l}
\left(\begin{array}{l}
\left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right) \\
2\left(\begin{array}{l} w_4^3\mu_2 - \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right) w_4^3\mu_3 \\ + \left(\sigma_{23}\sigma_{34}^{-(3,4)} + \sigma_{24}\sigma_{44}^{-(3,4)}\right)\left(w_4^4 - w_4^3\mu_4\right) \end{array}\right) \\
\left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right)
\end{array}\right) + \\
\left(\begin{array}{l}
\left(\begin{array}{l} w_4^3\mu_2 - \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right) w_4^3\mu_3 \\ + \left(\sigma_{23}\sigma_{34}^{-(3,4)} + \sigma_{24}\sigma_{44}^{-(3,4)}\right)\left(w_4^4 - w_4^3\mu_4\right) \end{array}\right) \\
\left(\sigma_{12}\sigma_{23}^{-(2,3,4)} + \sigma_{13}\sigma_{33}^{-(2,3,4)} + \sigma_{14}\sigma_{43}^{-(2,3,4)}\right)
\end{array}\right) + \\
\left(\begin{array}{l}
\left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right) \\
\left(\begin{array}{l} w_4^3\mu_1 \\ - \left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right) w_4^3\mu_2 \\ - \left(\sigma_{12}\sigma_{23}^{-(2,3,4)} + \sigma_{13}\sigma_{33}^{-(2,3,4)} + \sigma_{14}\sigma_{43}^{-(2,3,4)}\right) w_4^3\mu_3 \\ + \left(\sigma_{12}\sigma_{24}^{-(2,3,4)} + \sigma_{13}\sigma_{34}^{-(2,3,4)} + \sigma_{14}\sigma_{44}^{-(2,3,4)}\right)\left(w_4^4 - w_4^3\mu_4\right) \end{array}\right)
\end{array}\right)
\end{array}\right)
\end{aligned}
\tag{A.72}
$$

$$
\begin{aligned}
w_4 b = &\\
&\left(\begin{array}{l}
\left(\begin{array}{l}
\left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right) \\
2\left(\begin{array}{l} w_4^2\mu_2 - \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right) w_4^2\mu_3 \\ + \left(\sigma_{23}\sigma_{34}^{-(3,4)} + \sigma_{24}\sigma_{44}^{-(3,4)}\right)\left(w_4^3 - w_4^2\mu_4\right) \end{array}\right) \\
\left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right)
\end{array}\right) + \\
\left(\begin{array}{l}
\left(\begin{array}{l} w_4^2\mu_2 - \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right) w_4^2\mu_3 \\ + \left(\sigma_{23}\sigma_{34}^{-(3,4)} + \sigma_{24}\sigma_{44}^{-(3,4)}\right)\left(w_4^3 - w_4^2\mu_4\right) \end{array}\right) \\
\left(\sigma_{12}\sigma_{23}^{-(2,3,4)} + \sigma_{13}\sigma_{33}^{-(2,3,4)} + \sigma_{14}\sigma_{43}^{-(2,3,4)}\right)
\end{array}\right) + \\
\left(\begin{array}{l}
\left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right) \\
\left(\begin{array}{l}
w_4^2\mu_1 \\
- \left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right) w_4^2\mu_2 \\
- \left(\sigma_{12}\sigma_{23}^{-(2,3,4)} + \sigma_{13}\sigma_{33}^{-(2,3,4)} + \sigma_{14}\sigma_{43}^{-(2,3,4)}\right) w_4^2\mu_3 \\
+ \left(\sigma_{12}\sigma_{24}^{-(2,3,4)} + \sigma_{13}\sigma_{34}^{-(2,3,4)} + \sigma_{14}\sigma_{44}^{-(2,3,4)}\right)\left(w_4^3 - w_4^2\mu_4\right)
\end{array}\right)
\end{array}\right)
\end{array}\right)
\end{aligned}
\tag{A.73}
$$

Now we will begin calculating $\mathrm{E}\left[c\mathrm{E}\left[w_3|w_4\right]\right]$. We begin by simplifying c . The coefficient c , in (A.60) has two additive terms. We simplify the first additive term

now.

$$\begin{pmatrix} \left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right) w_4 \\ \left(\left(\sigma_{22} - \left(\begin{array}{l} \sigma_{23}\left(\sigma_{33}^{-(3,4)}\sigma_{32} + \sigma_{34}^{-(3,4)}\sigma_{42}\right) + \\ \sigma_{24}\left(\sigma_{43}^{-(3,4)}\sigma_{32} + \sigma_{44}^{-(3,4)}\sigma_{42}\right) \end{array} \right)\right) \right) \\ + \left(\begin{array}{l} \mu_2 - \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right)\mu_3 \\ + \left(\sigma_{23}\sigma_{34}^{-(3,4)} + \sigma_{24}\sigma_{44}^{-(3,4)}\right)(w_4 - \mu_4) \end{array} \right)^2 \end{pmatrix} = \\ \begin{pmatrix} \left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right) \\ \left(w_4 \left(\sigma_{22} - \left(\begin{array}{l} \sigma_{23}\left(\sigma_{33}^{-(3,4)}\sigma_{32} + \sigma_{34}^{-(3,4)}\sigma_{42}\right) + \\ \sigma_{24}\left(\sigma_{43}^{-(3,4)}\sigma_{32} + \sigma_{44}^{-(3,4)}\sigma_{42}\right) \end{array} \right)\right) \right) \\ + w_4 \left(\begin{array}{l} \left(\sigma_{23}\sigma_{34}^{-(3,4)} + \sigma_{24}\sigma_{44}^{-(3,4)}\right) w_4 + \\ \left(\begin{array}{l} \mu_2 - \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right)\mu_3 \\ -\mu_4\left(\sigma_{23}\sigma_{34}^{-(3,4)} + \sigma_{24}\sigma_{44}^{-(3,4)}\right) \end{array} \right) \end{array} \right)^2 \end{pmatrix} \tag{A.74}$$

Now we carry expand out the squared term and carry through multiplication of the outer w_4 term.

$$\begin{pmatrix} \left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right) \\ \left(w_4 \left(\sigma_{22} - \left(\begin{array}{l} \sigma_{23}\left(\sigma_{33}^{-(3,4)}\sigma_{32} + \sigma_{34}^{-(3,4)}\sigma_{42}\right) + \\ \sigma_{24}\left(\sigma_{43}^{-(3,4)}\sigma_{32} + \sigma_{44}^{-(3,4)}\sigma_{42}\right) \end{array} \right)\right) \right) \\ + \left(\sigma_{23}\sigma_{34}^{-(3,4)} + \sigma_{24}\sigma_{44}^{-(3,4)}\right)^2 w_4^3 + \\ \left(\begin{array}{l} 2\left(\sigma_{23}\sigma_{34}^{-(3,4)} + \sigma_{24}\sigma_{44}^{-(3,4)}\right) \\ \left(\begin{array}{l} \mu_2 - \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right)\mu_3 \\ -\mu_4\left(\sigma_{23}\sigma_{34}^{-(3,4)} + \sigma_{24}\sigma_{44}^{-(3,4)}\right) \end{array} \right) \end{array} \right) w_4^2 + \\ \left(\begin{array}{l} \mu_2 - \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right)\mu_3 \\ -\mu_4\left(\sigma_{23}\sigma_{34}^{-(3,4)} + \sigma_{24}\sigma_{44}^{-(3,4)}\right) \end{array} \right)^2 w_4 \end{pmatrix} \tag{A.75}$$

Now we will simplify the second additive term in c .

$$
\begin{pmatrix}
\begin{pmatrix}
\mu_2 - \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right)\mu_3 \\
+ \left(\sigma_{23}\sigma_{34}^{-(3,4)} + \sigma_{24}\sigma_{44}^{-(3,4)}\right)(w_4 - \mu_4)
\end{pmatrix} \\
\begin{pmatrix}
w_4\mu_1 \\
- \left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right) w_4\mu_2 \\
- \left(\sigma_{12}\sigma_{23}^{-(2,3,4)} + \sigma_{13}\sigma_{33}^{-(2,3,4)} + \sigma_{14}\sigma_{43}^{-(2,3,4)}\right) w_4\mu_3 \\
+ \left(\sigma_{12}\sigma_{24}^{-(2,3,4)} + \sigma_{13}\sigma_{34}^{-(2,3,4)} + \sigma_{14}\sigma_{44}^{-(2,3,4)}\right)(w_4^2 - w_4\mu_4)
\end{pmatrix}
\end{pmatrix} = \\
\begin{pmatrix}
\left(\mu_2 - \left(\sigma_{23}\sigma_{33}^{-(3,4)} + \sigma_{24}\sigma_{43}^{-(3,4)}\right)\mu_3 - \left(\sigma_{23}\sigma_{34}^{-(3,4)} + \sigma_{24}\sigma_{44}^{-(3,4)}\right)\mu_4\right) \\
\begin{pmatrix}
w_4\mu_1 \\
- \left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right) w_4\mu_2 \\
- \left(\sigma_{12}\sigma_{23}^{-(2,3,4)} + \sigma_{13}\sigma_{33}^{-(2,3,4)} + \sigma_{14}\sigma_{43}^{-(2,3,4)}\right) w_4\mu_3 \\
+ \left(\sigma_{12}\sigma_{24}^{-(2,3,4)} + \sigma_{13}\sigma_{34}^{-(2,3,4)} + \sigma_{14}\sigma_{44}^{-(2,3,4)}\right)(w_4^2 - w_4\mu_4)
\end{pmatrix}
\end{pmatrix} + \\
\begin{pmatrix}
\left(\sigma_{23}\sigma_{34}^{-(3,4)} + \sigma_{24}\sigma_{44}^{-(3,4)}\right) \\
\begin{pmatrix}
w_4^2\mu_1 \\
- \left(\sigma_{12}\sigma_{22}^{-(2,3,4)} + \sigma_{13}\sigma_{32}^{-(2,3,4)} + \sigma_{14}\sigma_{42}^{-(2,3,4)}\right) w_4^2\mu_2 \\
- \left(\sigma_{12}\sigma_{23}^{-(2,3,4)} + \sigma_{13}\sigma_{33}^{-(2,3,4)} + \sigma_{14}\sigma_{43}^{-(2,3,4)}\right) w_4^2\mu_3 \\
+ \left(\sigma_{12}\sigma_{24}^{-(2,3,4)} + \sigma_{13}\sigma_{34}^{-(2,3,4)} + \sigma_{14}\sigma_{44}^{-(2,3,4)}\right)(w_4^3 - w_4^2\mu_4)
\end{pmatrix}
\end{pmatrix}
\end{aligned} \tag{A.76}
$$

Now we have simplified c. It is time to take the expectation of $c\mathrm{E}\left[w_3|w_4\right]$.

$$
\begin{aligned}
&\mathrm{E}\left[c\mathrm{E}\left[w_3|w_4\right]\right] = \mathrm{E}\left[c\left(\mu_3 + \tfrac{\sigma_{34}}{\sigma_{44}}\left(w_4 - \mu_4\right)\right)\right] = \\
&\mathrm{E}\left[c\left(\mu_3 - \tfrac{\sigma_{34}}{\sigma_{44}}\mu_4\right) + \tfrac{\sigma_{34}}{\sigma_{44}}cw_4\right] = \\
&\left(\mu_3 - \tfrac{\sigma_{34}}{\sigma_{44}}\mu_4\right)\mathrm{E}\left[c\right] + \tfrac{\sigma_{34}}{\sigma_{44}}\mathrm{E}\left[cw_4\right]
\end{aligned} \tag{A.77}
$$

The expectation $\mathrm{E}\left[c\right]$ is easily calculated using (A.11). Similarly,$\mathrm{E}\left[cw_4\right]$ is easily calculated using (A.11) and (A.10). To demonstrate this, we will perform the outer multiplication in cw_4. This involves multiplication of the first additive term of c,

(A.75) by w_4.

$$\left(\begin{array}{l}\left(\sigma_{12}\sigma_{22}^{-(2,3,4)}+\sigma_{13}\sigma_{32}^{-(2,3,4)}+\sigma_{14}\sigma_{42}^{-(2,3,4)}\right)\\ \left(\begin{array}{l} w_4^2\left(\sigma_{22}-\left(\begin{array}{l}\sigma_{23}\left(\sigma_{33}^{-(3,4)}\sigma_{32}+\sigma_{34}^{-(3,4)}\sigma_{42}\right)+\\ \sigma_{24}\left(\sigma_{43}^{-(3,4)}\sigma_{32}+\sigma_{44}^{-(3,4)}\sigma_{42}\right)\end{array}\right)\right)\\ +\left(\sigma_{23}\sigma_{34}^{-(3,4)}+\sigma_{24}\sigma_{44}^{-(3,4)}\right)^2 w_4^4+\\ \left(\begin{array}{l}2\left(\sigma_{23}\sigma_{34}^{-(3,4)}+\sigma_{24}\sigma_{44}^{-(3,4)}\right)\\ \left(\begin{array}{l}\mu_2-\left(\sigma_{23}\sigma_{33}^{-(3,4)}+\sigma_{24}\sigma_{43}^{-(3,4)}\right)\mu_3\\ -\mu_4\left(\sigma_{23}\sigma_{34}^{-(3,4)}+\sigma_{24}\sigma_{44}^{-(3,4)}\right)\end{array}\right)\end{array}\right) w_4^3+\\ \left(\begin{array}{l}\mu_2-\left(\sigma_{23}\sigma_{33}^{-(3,4)}+\sigma_{24}\sigma_{43}^{-(3,4)}\right)\mu_3\\ -\mu_4\left(\sigma_{23}\sigma_{34}^{-(3,4)}+\sigma_{24}\sigma_{44}^{-(3,4)}\right)\end{array}\right)^2 w_4^2\end{array}\right)\end{array}\right) \tag{A.78}$$

This term is added to the multiplication of the second additive term (A.76), by w_4.

$$\left(\begin{array}{l}\left(\mu_2-\left(\sigma_{23}\sigma_{33}^{-(3,4)}+\sigma_{24}\sigma_{43}^{-(3,4)}\right)\mu_3-\left(\sigma_{23}\sigma_{34}^{-(3,4)}+\sigma_{24}\sigma_{44}^{-(3,4)}\right)\mu_4\right)\\ \left(\begin{array}{l} w_4^2\mu_1\\ -\left(\sigma_{12}\sigma_{22}^{-(2,3,4)}+\sigma_{13}\sigma_{32}^{-(2,3,4)}+\sigma_{14}\sigma_{42}^{-(2,3,4)}\right)w_4^2\mu_2\\ -\left(\sigma_{12}\sigma_{23}^{-(2,3,4)}+\sigma_{13}\sigma_{33}^{-(2,3,4)}+\sigma_{14}\sigma_{43}^{-(2,3,4)}\right)w_4^2\mu_3\\ +\left(\sigma_{12}\sigma_{24}^{-(2,3,4)}+\sigma_{13}\sigma_{34}^{-(2,3,4)}+\sigma_{14}\sigma_{44}^{-(2,3,4)}\right)(w_4^3-w_4^2\mu_4)\end{array}\right)\end{array}\right)+$$
$$\left(\begin{array}{l}\left(\sigma_{23}\sigma_{34}^{-(3,4)}+\sigma_{24}\sigma_{44}^{-(3,4)}\right)\\ \left(\begin{array}{l} w_4^3\mu_1\\ -\left(\sigma_{12}\sigma_{22}^{-(2,3,4)}+\sigma_{13}\sigma_{32}^{-(2,3,4)}+\sigma_{14}\sigma_{42}^{-(2,3,4)}\right)w_4^3\mu_2\\ -\left(\sigma_{12}\sigma_{23}^{-(2,3,4)}+\sigma_{13}\sigma_{33}^{-(2,3,4)}+\sigma_{14}\sigma_{43}^{-(2,3,4)}\right)w_4^3\mu_3\\ +\left(\sigma_{12}\sigma_{24}^{-(2,3,4)}+\sigma_{13}\sigma_{34}^{-(2,3,4)}+\sigma_{14}\sigma_{44}^{-(2,3,4)}\right)(w_4^4-w_4^3\mu_4)\end{array}\right)\end{array}\right) \tag{A.79}$$

This completes the calculation of $\mathrm{E}\left[w_1w_2w_3w_4\right]$. Now we have all the moment

definitions that we need. We will summarize our calculations and then move forward.

$$\begin{aligned}
&\mathrm{E}\left[w_4\right] = (\text{A.5}) \\
&\mathrm{E}\left[w_4^2\right] = (\text{A.6}) \\
&\mathrm{E}\left[w_4 w_3\right] = (\text{A.7}) \\
&\mathrm{E}\left[w_4^4\right] = (\text{A.10}) \\
&\mathrm{E}\left[w_3^3 w_4\right] = (\text{A.19}) \\
&\mathrm{E}\left[w_2 w_3 w_4^2\right] = (\text{A.32}), (\text{A.10}), (\text{A.11}) \\
&\mathrm{E}\left[w_1 w_2 w_3 w_4\right] = (\text{A.42}) \text{ onward} \\
&\mathrm{E}\left[w_4^3\right] = (\text{A.11}) \\
&\mathrm{E}\left[w_4^2 w_3\right] = (\text{A.41}) \\
&\mathrm{E}\left[w_1 w_2 w_3\right] = (\text{A.40}), (\text{A.11})
\end{aligned} \tag{A.80}$$

VITA

Charles David Lindsey received his B.A. in mathematics and computer science in May 2002 from Southwestern University, where he also minored in philosophy. In December 2007, he received his M.S. in statistics from Texas A&M University, and his Ph.D. in statistics at Texas A&M University was awarded in May 2010. His current reseach interests are multivariate and non-parametric statistics. His address is: Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX 77843-3143.

The typist for this thesis was Charles Lindsey.